**Structural Biology**
**Prof. Saugata Hazra**
**Department of Biotechnology**
**Indian Institute of Technology – Roorkee**

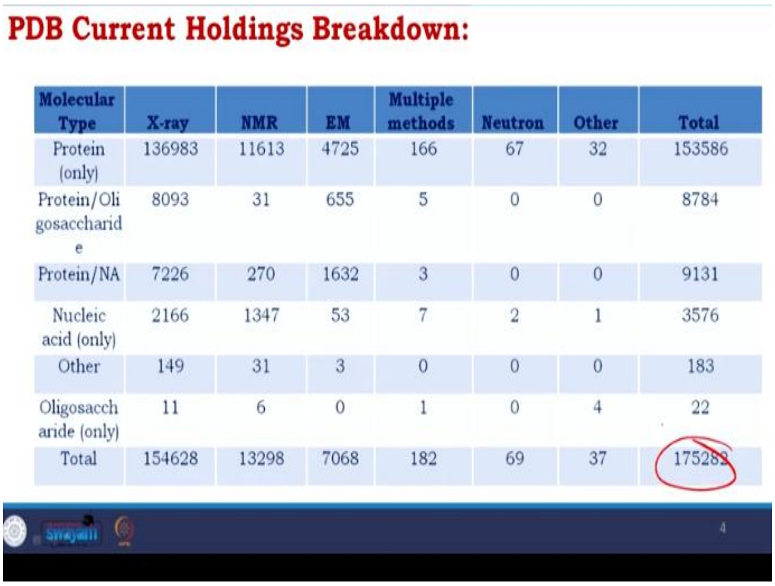
# Lecture – 42
# Description of Structure-Related Files (.pdb, .mmcif, .mtz, etcetera.)

Hi everyone, welcome again to the course of structural biology, we are going through the module on visualization, I have given you the details about the journey of how modelling visualizations started with the initial development of the physical model to the sophistication of the model to the sculpturing and ultimately to the introduction of the computer. And today it really has achieved a place where very, very high-resolution pictures are available.

High-resolution movies are available, which are helping scientists, educators, researchers, and even students to understand biology in an easy mode. But there is a process that, we cannot talk to the computer. So, today we will talk about that today we will talk about the database of structures or biological macromolecules. In the last module, I have already talked about the Protein Data Bank, which is considered the de database of 3D structure.

But in that database, how the files are there, and what types of files are there? And especially the coordinate files which are the main to give us the 3D information of the atoms, we will understand that in this class. So, I have talked about protein data banks. I have also talked about how different databases, the NMR database, the cryo-electron database, other databases like nucleotide database NDB and different parts like PDBE the PDB from Europe, the PDBJ the PDB from Japan, all come into a common database or data bank which is called protein databank RCSB PDB.

**(Refer Slide Time: 02:51)**



PDB Current Holdings Breakdown:

| Molecular Type | X-ray | NMR | EM | Multiple methods | Neutron | Other | Total |
|---|---|---|---|---|---|---|---|
| Protein (only) | 136983 | 11613 | 4725 | 166 | 67 | 32 | 153586 |
| Protein/Oligosaccharide | 8093 | 31 | 655 | 5 | 0 | 0 | 8784 |
| Protein/NA | 7226 | 270 | 1632 | 3 | 0 | 0 | 9131 |
| Nucleic acid (only) | 2166 | 1347 | 53 | 7 | 2 | 1 | 3576 |
| Other | 149 | 31 | 3 | 0 | 0 | 0 | 183 |
| Oligosaccharide (only) | 11 | 6 | 0 | 1 | 0 | 4 | 22 |
| Total | 154628 | 13298 | 7068 | 182 | 69 | 37 | 175282 |

If you look at the current holdings of the protein databank, only protein entries are 136983 by solving through X-Ray structure, NMR it is 11613, electron microscopy 4725 and other multiple methods 166, neutron diffraction 67, other 32 which are coming into small methods. So, a total of 153586 protein molecules already have entered this structure. Protein oligosaccharide 8093 for X-Ray 31 for NMR, EM 655 is a significant change.

Protein nucleic acid is 7226 in X-Ray, 270 NMR, and a very significant number 1632 in electron microscopy this field this protein-nucleic acid because these are talking about big complexes ribosome is one of them and one day the electron microscope is going to surpass. Anyway, we have already talked about 175282 entries are there. But before going into detail about the PDB I want to talk about a few things which are very important.

As our journey started from sequence to function and for that structure is required. So, when we see here, 153586 proteins, only protein i.e., 153586. Do they represent that many proteins do that structural entry of 153586 proteins represent 153586 proteins? The answer is no, why not? Because many proteins are important because, of therapeutic reasons because of other reason like their antibody or their important structural protein and all these things.

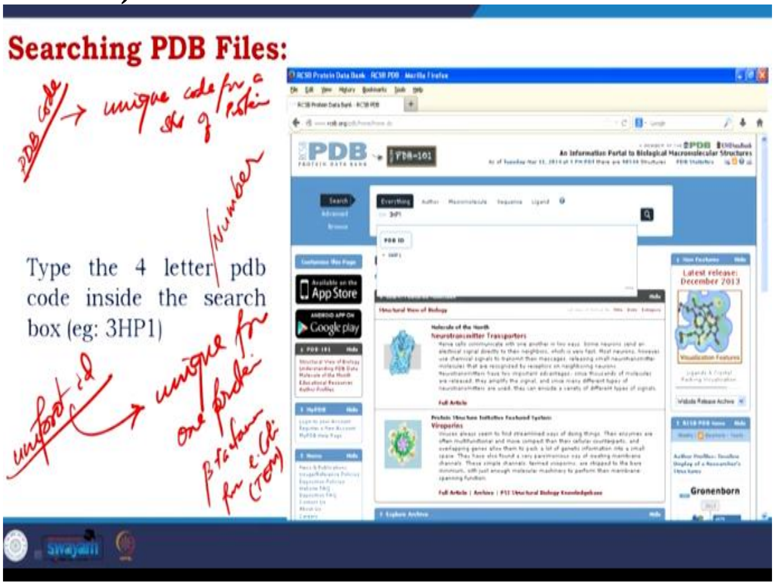
So, kind of 30,000 I would say unique proteins would be available so, this is a knowledge, another knowledge initially when the journey of structure determination was started, for protein, you know when we say the structure to function, the important thing is folds, how the protein is folded and that is why it is critical to go from a journey of a sequence to structure. Now, you would be very surprised or very excited to see that the appearance of new folds reduces with the increased appearance of new structures.

And currently, every year if you see, we have no new fold or very little modification in the existing fold or a permutation combination of 2 or 3 folds coming into one structure. So, this is a great thing because, with the next generation sequencing, we now know that every day 1000’s and millions of genes are appearing, which means millions of proteins, but if they correspond to just a few folds, then the predication could be easier our computational tools would be more accurate.

And the requirement of structure determination would become much less keep in mind that this would not reduce the importance of structural biologists, because you meet new structures, because of different reasons, like when you want to do a drug designing, you need to know the high-resolution interaction between the protein and drug, and you need crystallographers to work on them.

If you want to go for the high-resolution structure of the bigger complexes, you need electron microscopy, when you need the dynamics of the small protein’s peptides, you need NMR but with a smaller number of nuchal folds coming the accuracy of computational prediction will be enhanced and that would help us controlling and determining more and more structured through computational prediction.

**(Refer Slide Time: 08:25)**



## Searching PDB Files:

Type the 4 letter/number pdb code inside the search box (eg: 3HP1)

PDB code → unique code for a str of protein

uniprot id → unique for one protein

β galactosidase for E.Coli (TEM)

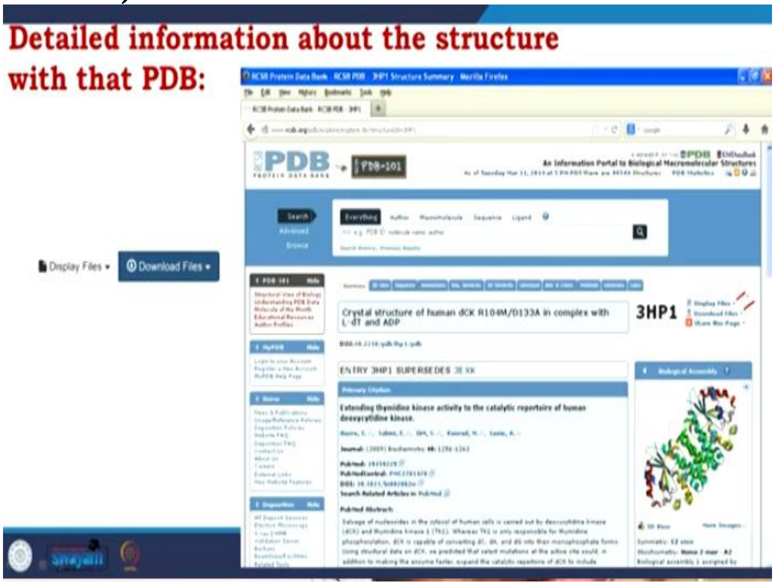
RCSB Protein Data Bank - RCSB PDB - Mozilla Firefox
PDB
PDB-101
An Information Portal to Biological Macromolecular Structures
Search
Advanced
Browse
Everything
Author
Macromolecule
Sequence
Ligand
3HP1
PDB ID
App Store
Google play
Neurotransmitter Transporters
Viroporins
Latest release: December 2013
Gronenborn
swayam

So, first, you need to get the PDB file, this is the homepage of PDB first you must search PDB file and for that, if you remember, you must use a 4-letter number combination, which is called a PDB code. I would like to repress your knowledge about a PDB code that gives a unique code for a structure of a protein which means one protein could have many PDB IDs, if that protein has more entries here, why it would need more entries?

Remember, just in the last slide I talked about if a protein is important, like a protein I work on which is called beta-lactamase. Beta-lactamase is a very important therapeutic protein. So, people need to understand beta-lactamase. So, they have tried to get its structure from different organisms. They tried to get the structure with different ligands drugs. And currently, you will get 1000's structures no single 1000's structures of beta-lactamase.

So, Unicode is PDB ID, but there is something called UniProt id, this is unique for one protein. So, if you have beta-lactamase from E. coli, then it would have only one UniProt ID, but it could have many PDB IDs, you have to understand that, so incorporate the unique PDB ID which is unique for a structured give search and then you will so here you put the PDB ID, I have given search for 3HP1.

**(Refer Slide Time: 10:56)**

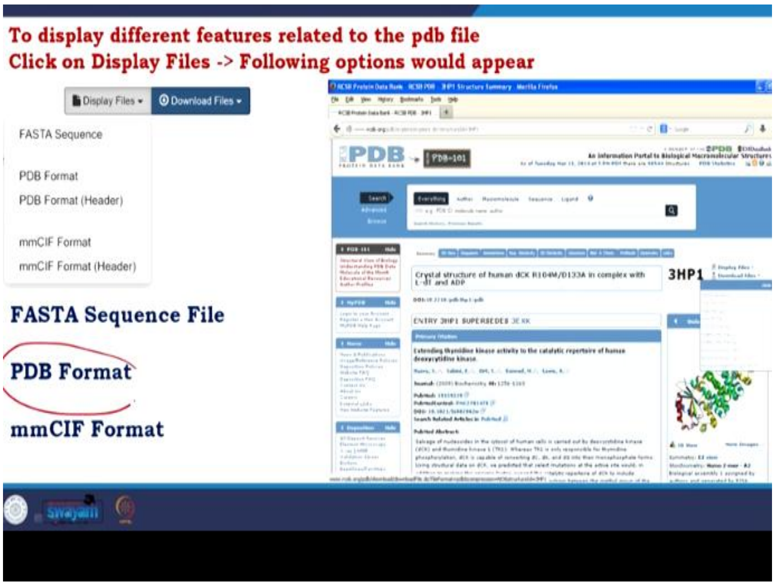


Detailed information about the structure with that PDB:

Display Files ▾
Download Files ▾
RCSB Protein Data Bank - RCSB PDB - 3HP1 Structure Summary - Mozilla Firefox
PDB
PROTEIN DATA BANK
PDB-101
An Information Portal to Biological Macromolecular Structures
Search
Advanced
Browse
Everything
Author
Macromolecule
Sequence
Ligand
Crystal structure of human dCK R104M/D133A in complex with L-dT and ADP
3HP1
Display Files
Download Files
Share this Page
ENTRY 3HP1 SUPERSEDES 3EXK
Primary Citation
Extending thymidine kinase activity to the catalytic repertoire of human deoxycytidine kinase.
Biological Assembly

And you will come to the information, this is the crystal structure of human DCK R104M, which is the information of a mutant, D 133 A another information of a mutant in complex with L-dT means L-deoxy Thymidine and ADP Adenosine Dinucleotide Phosphate. So, you see the entry and there are 2 options to look at the file. One option is to display files another option is to download files. To display files, you could see the files to download files, you would download the files.

**(Refer Slide Time: 11:45)**



To display different features related to the pdb file
Click on Display Files -> Following options would appear

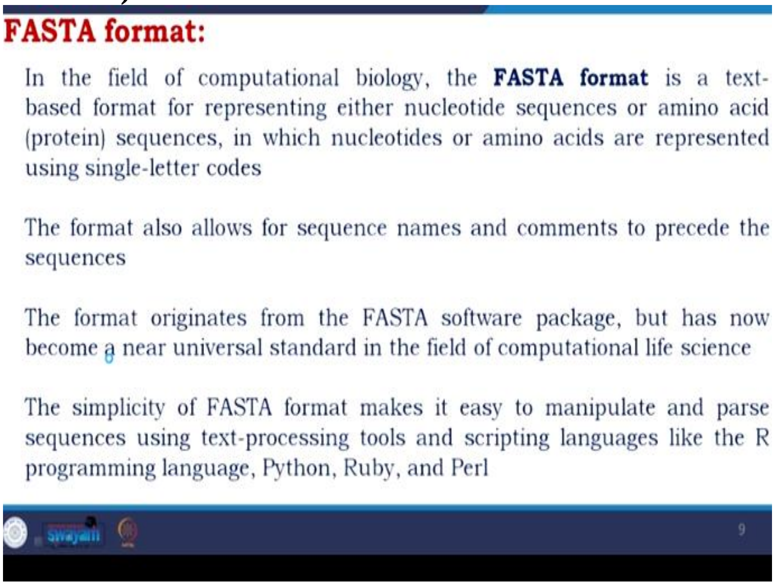
FASTA Sequence File

PDB Format

mmCIF Format

When you go to display files, there are 3 major groups of information the FASTA sequence, 2 the PDB format which is into us PDB format, the PDB format header, and the mmCIF format. PDB format I talked about, and I would discuss it in detail at the end of the class. So, now, I will discuss first a sequence file and mmCIF format.

**(Refer Slide Time: 12:18)**



## FASTA format:

In the field of computational biology, the **FASTA format** is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes

The format also allows for sequence names and comments to precede the sequences

The format originates from the FASTA software package, but has now become a near universal standard in the field of computational life science

The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages like the R programming language, Python, Ruby, and Perl

FASTA format In the field of computational biology, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid sequences in which nucleotides or amino acids are represented using single-letter codes. So, FASTA format is a format, which is a globally used standard format to write your protein or nucleic acid sequence.

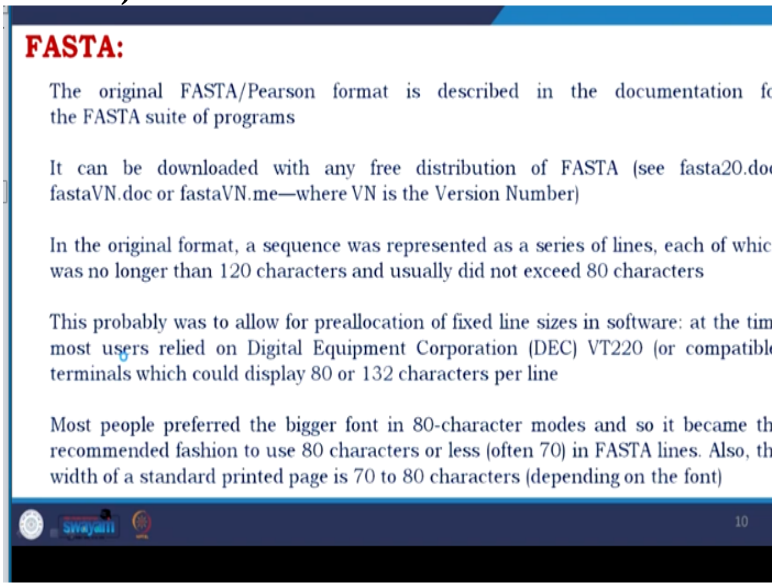
The format also allows for sequence names and comments to precede the sequences. The format originates from the FASTA software package but has now become a near-universal standard in the field of computational life science. As I told first time is good because we all use FASTA. So, any of the formats, you know when you talk about these, you should remember any format you cannot say this the good or bad, if a lot of people around the world start using them that format it would become a universal format.

And it would be easy to talk in that language, you know we talk in different languages, but it is a universal normal kind of when we meet worldwide people who speak in English, it is the same as you know in international unions have decided rules how to talk, how to write a compound globally IUPC. The IUPC rules decide what you call a compound you call it Ram, Sam, you call it they carry you call it anything.

But there should be a universal limb that everyone recognizes. FASTA format is kind of becoming a universal format representing the sequences. What is the reason? The simplicity of the FASTA format makes it easy to manipulate and parse sequences. Parse sequence means parse is a kind of you could say program which extracts information, so when you want to extract sequence information and the information is in FASTA format it is easy.

So, the simplicity of the FASTA format makes it easy to and parse sequences using text processing tools and scripting languages like our programming language Python, Ruby, Perl, all the commonly used languages.

**(Refer Slide Time: 15:10)**



**FASTA:**

The original FASTA/Pearson format is described in the documentation fc the FASTA suite of programs

It can be downloaded with any free distribution of FASTA (see fasta20.do fastaVN.doc or fastaVN.me—where VN is the Version Number)

In the original format, a sequence was represented as a series of lines, each of whic was no longer than 120 characters and usually did not exceed 80 characters

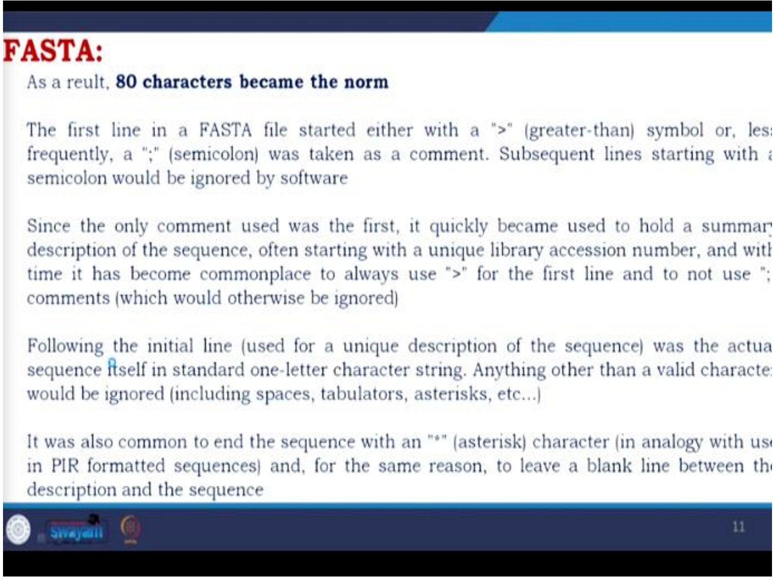
This probably was to allow for preallocation of fixed line sizes in software: at the tim most users relied on Digital Equipment Corporation (DEC) VT220 (or compatibl terminals which could display 80 or 132 characters per line

Most people preferred the bigger font in 80-character modes and so it became th recommended fashion to use 80 characters or less (often 70) in FASTA lines. Also, th width of a standard printed page is 70 to 80 characters (depending on the font)

The original FASTA or Pearson format, which is also alternatively called Pearson format is described in the documentation for the fastest suite of programs. You could download it with any free distribution of FASTA the format and all I have written for you in the original format is sequence was represented as a series of lines, each of which was no longer than 120 characters and usually did not exceed 80 characters. This probably was to allow for the pre-allocation of fixed line sizes in software.

At the time, most users relied on Digital Equipment Corporation or DEC VT220 terminals which could display 80 or 132 characters per line. So, to match with them, it had no longer than 120 characters and usually not exceeding 80 characters. Most people prefer the bigger font in 80-character modes and so it becomes the recommended fashion to use 80 characters or less often 17 FASTA lines also, the width of standard printed pages is 70 to 80 characters which also helps.

**(Refer Slide Time: 16:27)**



FASTA:

As a reult, **80 characters became the norm**

The first line in a FASTA file started either with a ">" (greater-than) symbol or, les frequently, a ";" (semicolon) was taken as a comment. Subsequent lines starting with semicolon would be ignored by software

Since the only comment used was the first, it quickly became used to hold a summar description of the sequence, often starting with a unique library accession number, and wit time it has become commonplace to always use ">" for the first line and to not use "; comments (which would otherwise be ignored)

Following the initial line (used for a unique description of the sequence) was the actua sequence itself in standard one-letter character string. Anything other than a valid characte would be ignored (including spaces, tabulators, asterisks, etc...)

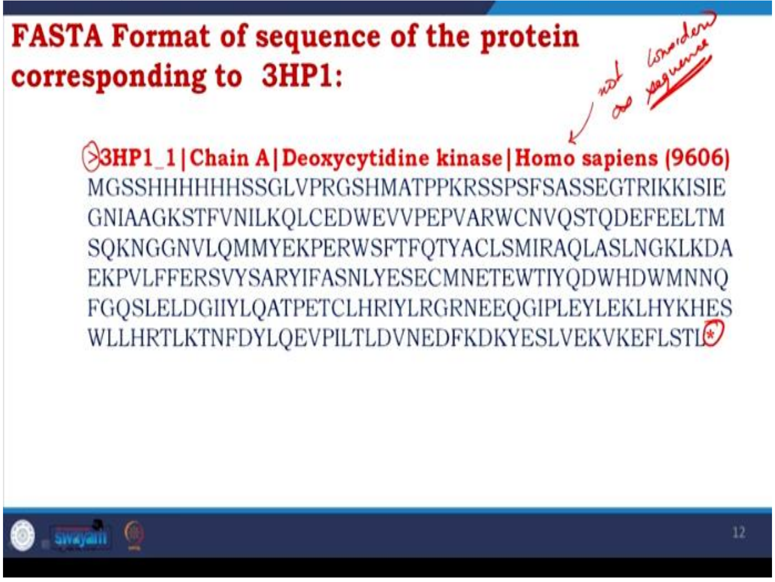
It was also common to end the sequence with an "*" (asterisk) character (in analogy with us in PIR formatted sequences) and, for the same reason, to leave a blank line between th description and the sequence

So, as a result, the character becomes a global norm. The first line in the FASTA file started with a greater than a symbol or sometimes a semicolon was taken as a comment, subsequent lines starting with his semi-colon would be ignored by the software. So, if you want to write some information for your understanding regarding the sequence, if you do this, then the software would not read them as a part of the sequence.

Since the only comment used was the first it quickly became used to hold a summary description of the sequence, often starting with a unique library access number, and with time it has become commonplace to always use greater than for the first line and to not use semicolon comments. Following the initial line used for a unique descriptor of the sequence was the actual sequence itself in standard letter, character string.

Anything other than a valid character would be ignored including spaces, tabulators, and asterisks. It was also common to end the sequence with a star asterisk character and for the same reason to leave a blank line between the description and the sequence.

**(Refer Slide Time: 17:51)**



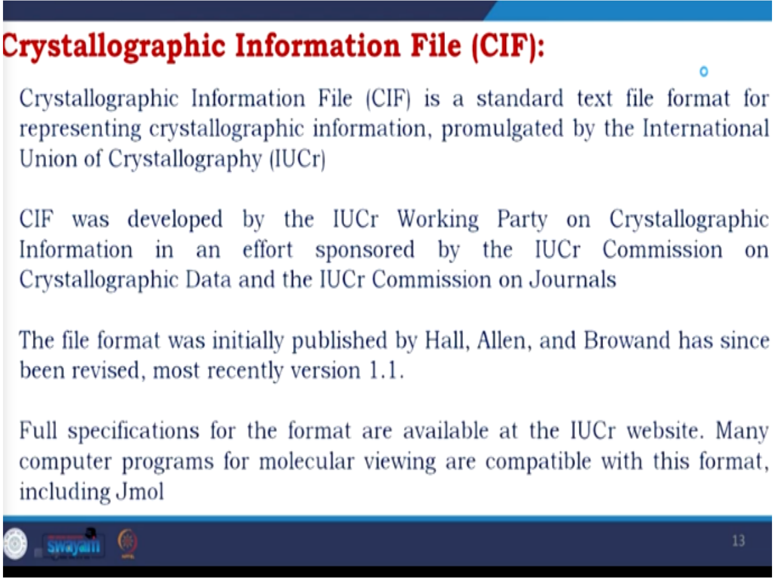
FASTA Format of sequence of the protein corresponding to 3HP1:

not considered as sequence

>3HP1_1|Chain A|Deoxycytidine kinase|Homo sapiens (9606)
MGSSHHHHHHSSGLVPRGSHMATPPKRSSPSFSASSEGTRIKKISIE
GNIAAGKSTFVNILKQLCEDWEVVPEPVARWCNVQSTQDEFEELTM
SQKNGGNVLQMMYEKPERWSFTFQTYACLSMIRAQLASLNGKLKDA
EKPVLFFERSVYSARYIFASNLYESECMNETEWTIYQDWHDWMNNQ
FGQSLELDGIIYLQATPETCLHRIYLRGRNEEQGIPLEYLEKLHYKHES
WLLHRTLKTNFDYLQEVPILTLDVNEDFKDKYESLVEKVKEFLSTL*

This is the FASTA format representation of the protein corresponding to the PDB ID 3HP1. If you see it starts with it greater than sign it ends with an asterisk sign and the description is here, this line is not considered as part of the sequence because of the greater than you would get up then you could write more things here. The next sequences are giving amino acid sequence AMG and all these things. So, M is methanine, G is for glycine and you get the entire sequence of protein which is used for structure determination.

**(Refer Slide Time: 18:40)**



Crystallographic Information File (CIF):

Crystallographic Information File (CIF) is a standard text file format for representing crystallographic information, promulgated by the International Union of Crystallography (IUCr)

CIF was developed by the IUCr Working Party on Crystallographic Information in an effort sponsored by the IUCr Commission on Crystallographic Data and the IUCr Commission on Journals

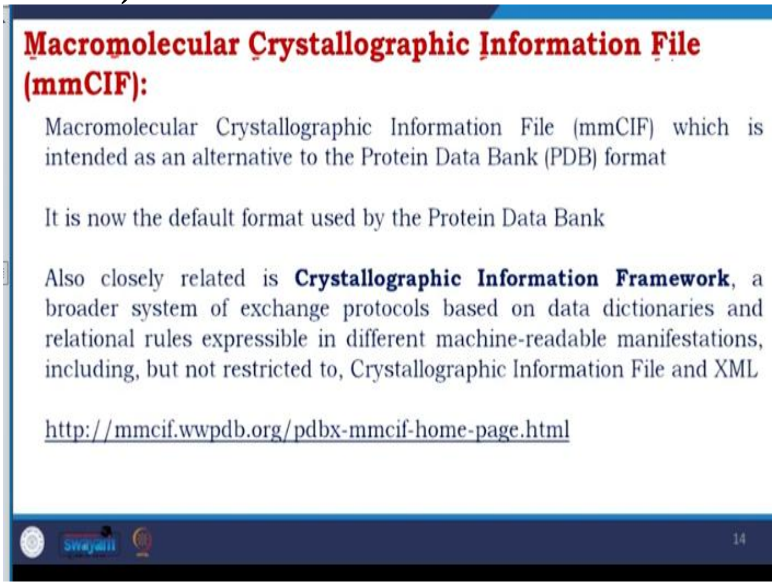
The file format was initially published by Hall, Allen, and Browand has since been revised, most recently version 1.1.

Full specifications for the format are available at the IUCr website. Many computer programs for molecular viewing are compatible with this format, including Jmol

Coming to the second one the mmCIF but before describing mmCIF I would describe the CIF file, CIF piles are more popular, and mostly used in any crystal information so it is called the crystallographic information file of CIF. The crystallographic information file is a standard text file format for representing crystallographic information, promulgated by the International Union of Crystallography the IUCr. As I told all those standard files which are used globally, have some universal rules behind them.

And the crystallography information file the rules are made by the International Union of Crystallography the ICUr. CIF was developed by the IUCr working party on crystallographic information in an effort sponsored by the ICUr Commission on Crystallographic Data and the IUCr Commission on Journals. The file format was initially published by Hall, Allen, and Browand has since been revised and the most recent version is 1.1. Full specifications for the format are available on the ICUr website. Many computer programs per molecular views are compatible with this format including Jmol.

**(Refer Slide Time: 20:06)**



## Macromolecular Crystallographic Information File (mmCIF):

Macromolecular Crystallographic Information File (mmCIF) which is intended as an alternative to the Protein Data Bank (PDB) format

It is now the default format used by the Protein Data Bank

Also closely related is **Crystallographic Information Framework**, a broader system of exchange protocols based on data dictionaries and relational rules expressible in different machine-readable manifestations, including, but not restricted to, Crystallographic Information File and XML

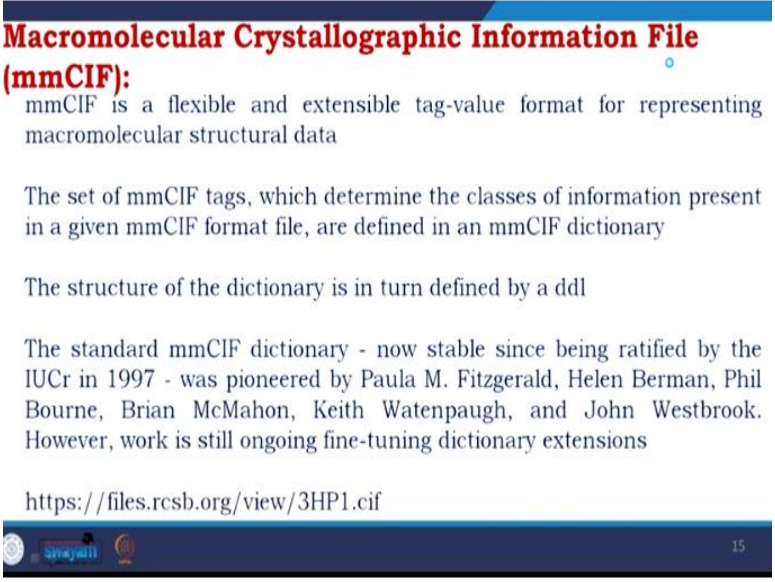
http://mmcif.wwpdb.org/pdbx-mmcif-home-page.html

14

mmCIF, which is our interest is Macromolecular Crystallographic Information File. So, the term mm is for macromolecular then for crystallographic information file, so mmCIF. mmCIF Macromolecular Crystallographic Information file which is intended as an alternative to the Protein Data Bank format. So, as I told we talk about PDB format, we are going to talk about the PDB format even in more detail, but mmCIF is an alternative to the Protein Data Bank format. It is now the default format used by Protein Data Bank.

Also closely related is the Crystallographic Information Framework, a broader system of exchange protocols based on data dictionaries and relational rules expressible in different missing readable manifestations including but not restricted to crystallographic information files and XML. So, there is an interconvertible option and that makes the use broader. This is the link where you could know more about the mmCIF.

**(Refer Slide Time: 21:29)**



**Macromolecular Crystallographic Information File (mmCIF):**

mmCIF is a flexible and extensible tag-value format for representing macromolecular structural data

The set of mmCIF tags, which determine the classes of information present in a given mmCIF format file, are defined in an mmCIF dictionary

The structure of the dictionary is in turn defined by a ddl

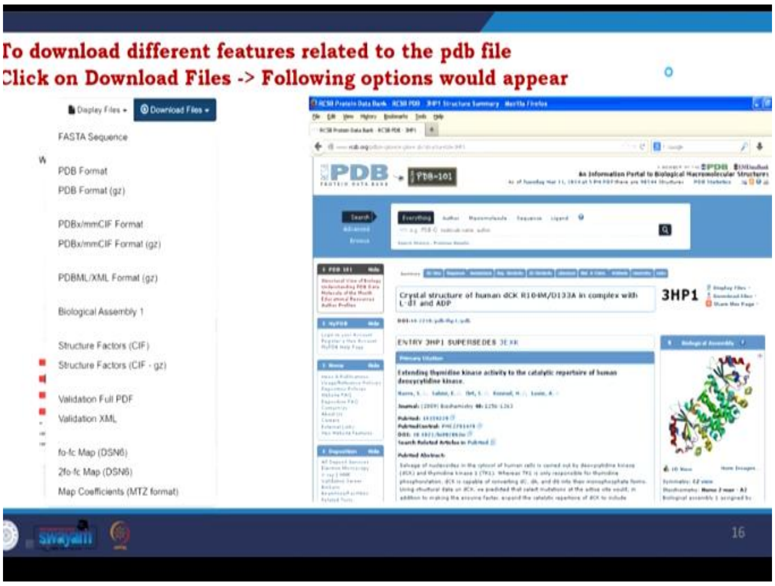
The standard mmCIF dictionary - now stable since being ratified by the IUCr in 1997 - was pioneered by Paula M. Fitzgerald, Helen Berman, Phil Bourne, Brian McMahon, Keith Watenpaugh, and John Westbrook. However, work is still ongoing fine-tuning dictionary extensions

https://files.rcsb.org/view/3HP1.cif

mmCIF is a flexible and extensible tag-value format for representing macromolecular structural data. The set of mmCIF tags which determine the classes of information present in a given mmCIF format file are defined in a mmCIF dictionary. The structure of the dictionary is in turn defined by DDL. The standard MMC dictionary now stable since being ratified by the ICUr in 1997 was pioneered by Paula M. Fitzgerald, Helen Berman, Phil Bourne, Brian McMahon, Keith, Watenpaugh, and John Westbrook.

However, what is still going to improve this is the link of the mmCIF file of that protein whose unique structure is provided by the PDB ID 3HP1.

**(Refer Slide Time: 22:28)**



To download different features related to the pdb file
Click on Download Files -> Following options would appear

- Display Files
- Download Files
- 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)

Crystal structure of human dCK R104M/D133A in complex with L-dT and ADP

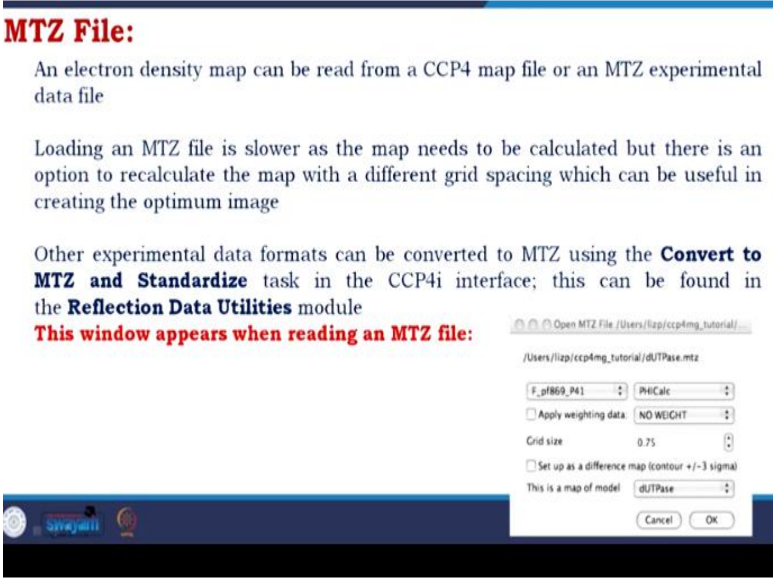
3HP1

ENTRY 3HP1 SUPERSEDES 3EXK

Coming to the downloading, so there are 2 options display file download file, the download file option is more vivid here we have first a sequence format I have already talked about PDB format we have discussed PDBx/mmCIF, and we have discussed PDBML/XML another alternative biological assembly. So, if the protein is oligomeric when you look at the coordinate, you see the file but the actual biological assemblies expressed here.

Then we have the structural factors, we have validation, and we have maps, I will just talk about 2 things one is validation because that is very important and another is a map which I will discuss in the next class in detail when I would include the actual MTZ file and use it in the good format but here I will little bit introduce you too.

**(Refer Slide Time: 23:36)**



MTZ File:

An electron density map can be read from a CCP4 map file or an MTZ experimental data file

Loading an MTZ file is slower as the map needs to be calculated but there is an option to recalculate the map with a different grid spacing which can be useful in creating the optimum image

Other experimental data formats can be converted to MTZ using the **Convert to MTZ and Standardize** task in the CCP4i interface; this can be found in the **Reflection Data Utilities** module

This window appears when reading an MTZ file:

Open MTZ File /Users/lizp/ccp4mg_tutorial/...

/Users/lizp/ccp4mg_tutorial/dUTPase.mtz

F_pf869_P41 | PHICalc

□ Apply weighting data: NO WEIGHT

Grid size 0.75

□ Set up as a difference map (contour +/-3 sigma)

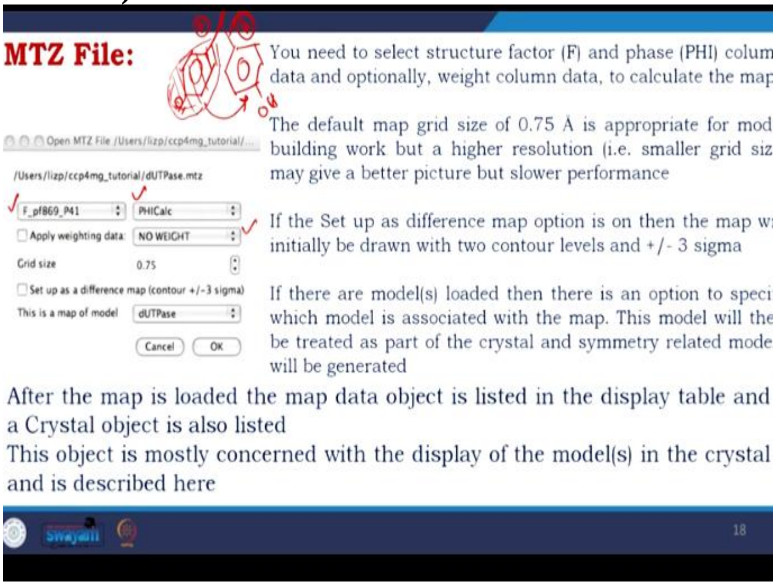
This is a map of model dUTPase

So, coming to the MTZ file, an electron density map can be read from the CCP4 map file or an MTZ experimental data file. So, up to now we are talking about PDB, PDB mmCIF, and

PDB XML they are all giving structures whereas MTZ is a different one that is giving the information about the map. If you go back, you will see if you remember the f0-fc map and 2f0-fc map, I have talked about them in the crystallography section.

Loading and MTZ file is slower as the map needs to be calculated but there is an option to recalculate the map with a different grid spacing which can be useful in creating the optimum image. Other experimental data formats can be converted to MTZ using that convert to MTZ and standardize tasks in the CCP4i interface in the CCP4i modules. This can be found in the reflection data utility module. This is the representation so if you look at this window, that is how the window appears when you would have to read a MTZ file.

**(Refer Slide Time: 24:50)**



## MTZ File:

○ ○ ○ Open MTZ File /Users/lizp/ccp4mg_tutorial/...

/Users/lizp/ccp4mg_tutorial/dUTPase.mtz

✓ F_pf869_P41 ✓ PHICalc

□ Apply weighting data: NO WEIGHT ✓

Grid size 0.75

□ Set up as a difference map (contour +/-3 sigma)

This is a map of model dUTPase

You need to select structure factor (F) and phase (PHI) colum data and optionally, weight column data, to calculate the map

The default map grid size of 0.75 A is appropriate for mod building work but a higher resolution (i.e. smaller grid siz may give a better picture but slower performance

If the Set up as difference map option is on then the map w initially be drawn with two contour levels and +/- 3 sigma

If there are model(s) loaded then there is an option to speci which model is associated with the map. This model will the be treated as part of the crystal and symmetry related mode will be generated

After the map is loaded the map data object is listed in the display table and a Crystal object is also listed

This object is mostly concerned with the display of the model(s) in the crystal and is described here

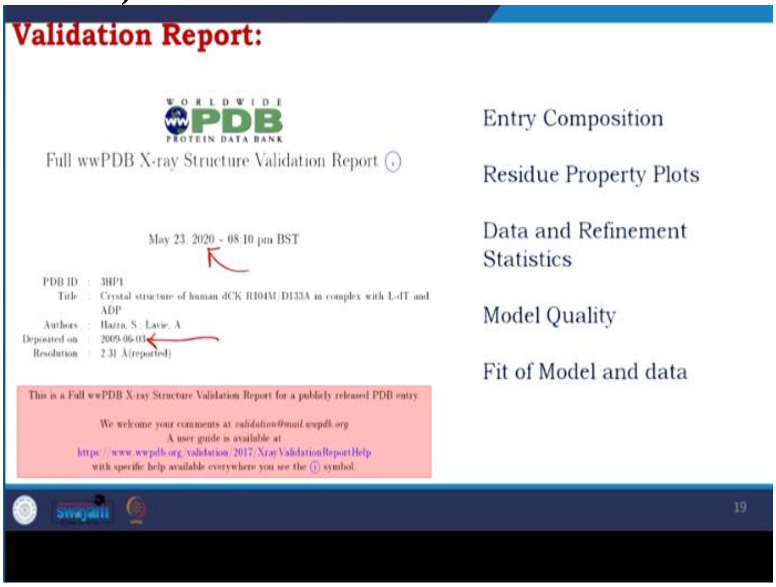
Going here you need to select structure factor F and the show structure factor and the phase column data and optionally, weight column data to calculate the map. So, weight column data to calculate the map, the default map grid size of 0.75 Angstrom is appropriate for model building work, but a higher resolution may give a better picture but slower performance because more high resolution more calculation.

If the setup as difference map option is on, then the map will initially be drawn with 2 contour levels with plus minus 3 sigmas. If there are models loaded then there is an option to specify which model is associated with the map, this model will then be treated as part of the crystal, and symmetry-related models will be generated. So, what it is talking about, is suppose you have this electron density, and you have 2PDB, this is a tyrosine suppose this is here and there is another tyrosine coming from another structure.

So, if you specify then the map will take proper space, and it might be shipped from here to here, if you specify which one is a structured is a structure of the, the structure, the structure which is assigned the model which is assigned for the map, the map will go and fit it so, that is what you could see. After the map is loaded, the map data object is listed in the display table and a crystal object is also listed.

This object is mostly concerned with the display of the model in that crystal and is described. So, when you look at that, you would be looking that the map would now guide you towards the further model building making the model complete. That is where the role of the MTZ files. The validation report is very important and incorporation of validation report the validation report was not there at the initial Protein Data Bank and when I say initial, it was incorporated a long time.

**(Refer Slide Time: 27:27)**



Validation Report:

Full wwPDB X-ray Structure Validation Report ⓘ

May 23, 2020 - 08:10 pm BST

| | | |
|---|---|---|
| PDB ID | : | 3HP1 |
| Title | : | Crystal structure of human dCK R104M/D133A in complex with L-dT and ADP |
| Authors | : | Hazra, S.; Lavie, A |
| Deposited on | : | 2009-06-03 |
| Resolution | : | 2.31 Å(reported) |

This is a Full wwPDB X-ray Structure Validation Report for a publicly released PDB entry.

We welcome your comments at *validation@mail.wwpdb.org*
A user guide is available at
https://www.wwpdb.org/validation/2017/XrayValidationReportHelp
with specific help available everywhere you see the ⓘ symbol.

Entry Composition

Residue Property Plots

Data and Refinement Statistics

Model Quality

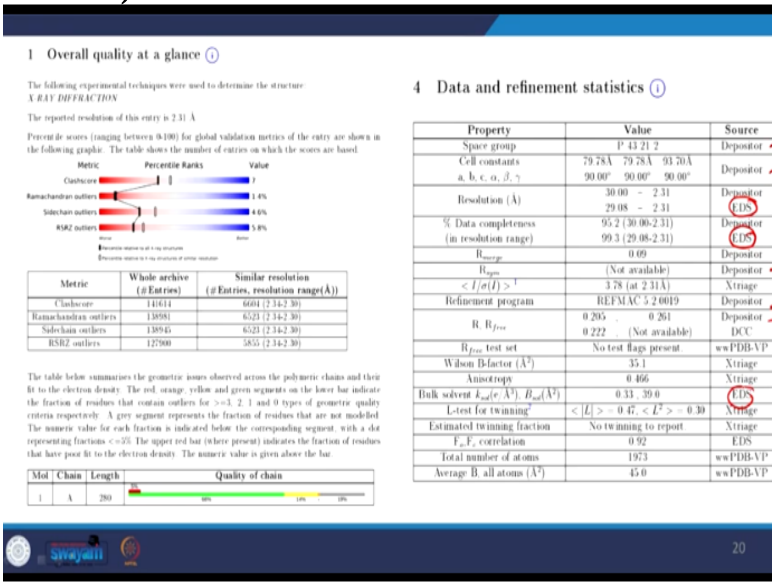
Fit of Model and data

So, if you look at the initial page of the validation report of 3HP1, you will see that the 3HP1 was deposited in 2009 but the validation report makes on May 23, 2020. So, now, probably you understand what I am saying, but the incorporation of the validation report came much later, but, these have changed a lot in the world of structural biology. Incorporation of a validation report has helped or guided the user towards the modification of the initial mistakes and when the user has taken care of that.

And when PDB agreed to the performance of the structural biologist and agreed to provide a validation report the quality of structure is significantly improved, which is further improving the quality of computational biology work. So, in one word, the introduction of validation reports is a revolution. There are many aspects of entry composition, residue property plots, data and refinement statistics, and model quality fit of model and data.

A lot of things happen I just keep this select this for giving you little detail on the entry composition where everything is there and data and refinement statistics, but the residue property plot works on individual residues how they are plotted, how they fit, they are in a grid position in terms of Ramachandran plot and all model qualities checking how the model quality is in comparison to other published structures and fit of model and data is fitting it with considering different macromolecules and all.

**(Refer Slide Time: 29:51)**



1 Overall quality at a glance

The following experimental techniques were used to determine the structure:
*X-RAY DIFFRACTION*

The reported resolution of this entry is 2.31 Å

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.

| Metric | Whole archive<br>(#Entries) | Similar resolution<br>(#Entries, resolution range(Å)) |
|---|---|---|
| Clashscore | 141614 | 6604 (2.34-2.30) |
| Ramachandran outliers | 138981 | 6523 (2.34-2.30) |
| Sidechain outliers | 138945 | 6523 (2.34-2.30) |
| RSRZ outliers | 127900 | 5855 (2.34-2.30) |

The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria respectively. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5% The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.

| Mol | Chain | Length | Quality of chain |
|---|---|---|---|
| 1 | A | 280 | 5% 68% 14% 19% |

4 Data and refinement statistics

| Property | Value | Source |
|---|---|---|
| Space group | P 43 21 2 | Depositor |
| Cell constants<br>a, b, c, α, β, γ | 79.78Å 79.78Å 93.70Å<br>90.00° 90.00° 90.00° | Depositor |
| Resolution (Å) | 30.00 - 2.31<br>29.08 - 2.31 | Depositor<br>EDS |
| % Data completeness<br>(in resolution range) | 95.2 (30.00-2.31)<br>99.3 (29.08-2.31) | Depositor<br>EDS |
| $R_{merge}$ | 0.09 | Depositor |
| $R_{sym}$ | (Not available) | Depositor |
| $< I/\sigma(I) >$[1] | 3.78 (at 2.31Å) | Xtriage |
| Refinement program | REFMAC 5.2.0019 | Depositor |
| R, $R_{free}$ | 0.205 , 0.261<br>0.222 , (Not available) | Depositor<br>DCC |
| $R_{free}$ test set | No test flags present. | wwPDB-VP |
| Wilson B-factor ($Å^2$) | 35.1 | Xtriage |
| Anisotropy | 0.466 | Xtriage |
| Bulk solvent $k_{sol}(e/Å^3)$, $B_{sol}(Å^2)$ | 0.33 , 39.0 | EDS |
| L-test for twinning[2] | $< \|L\| > = 0.47$, $< L^2 > = 0.30$ | Xtriage |
| Estimated twinning fraction | No twinning to report. | Xtriage |
| $F_o$,$F_c$ correlation | 0.92 | EDS |
| Total number of atoms | 1973 | wwPDB-VP |
| Average B, all atoms ($Å^2$) | 45.0 | wwPDB-VP |

20

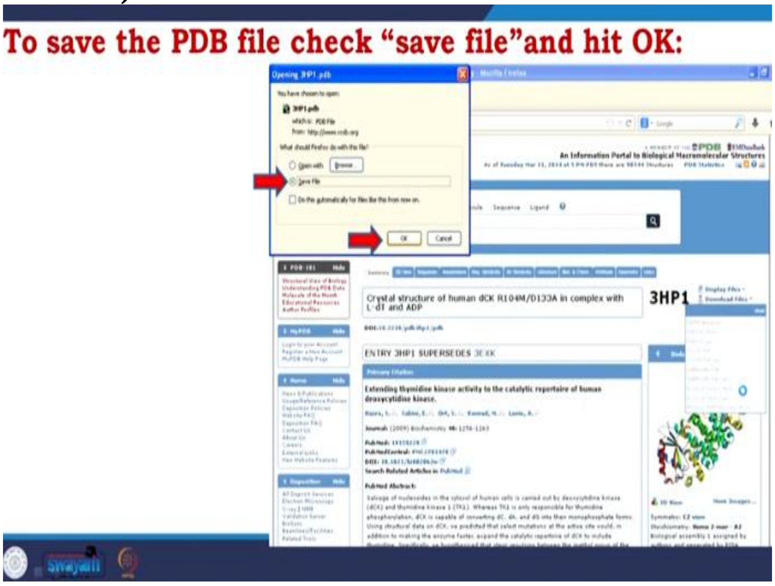
So, as I told overall quality at a glance here, the following experimental techniques are used to determine the structure extra diffraction. So, the technique is given the resolution is 2.31 the resolution is given the percentile score for the global validation matrix of the entry as shown in the following graphic. So, if you see where it is ranked, you see from worse to better. So, you will see the cohesion of the structure. Then in the matrix, you have class core Ramachandran outliers, sidechain outliers, and RSRZ outliers.

Also the table here, you could see the summary of geometric issues which are observed across the polymeric chains and they are fit to the electron density. So, where you see the green is better and you could see all the reasons for the colorings. Coming to data and refinements statistics here space groups, cell constants, resolution, percentage data completeness, R merge, R sim, signal to noise, I / I sigma, refinement program, R free, R free of test set.

Wilson B-factor, Antisotopy, Bulk solvent L- test for twinning, Estimated twining fraction, F0, Fe correlation, the total number of attempts, and average B factor is given some of them are coming from the depositor. So, I was the depositor of the structure. So, these are information coming from me but then some are also tested if you see by the electron density server and others are coming from either Electron Density Server or PDB or Xtriage.

So, day by day now, most of the information's getting better like sometimes the information is not good, not because the actual quality of the structure is bad. I am talking about the old structures because those factors are not taken care could be one of the reasons. And now when we are checking it for new structures, we check all the factors before even the final deposition and entry.

**(Refer Slide Time: 32:37)**



To save the PDB file check “save file”and hit OK:

Opening 3HP1.pdb

You have chosen to open:

3HP1.pdb

which is: PDB File

from: http://www.rcsb.org

What should Firefox do with this file?

○ Open with [Browse...]

● Save File

□ Do this automatically for files like this from now on.

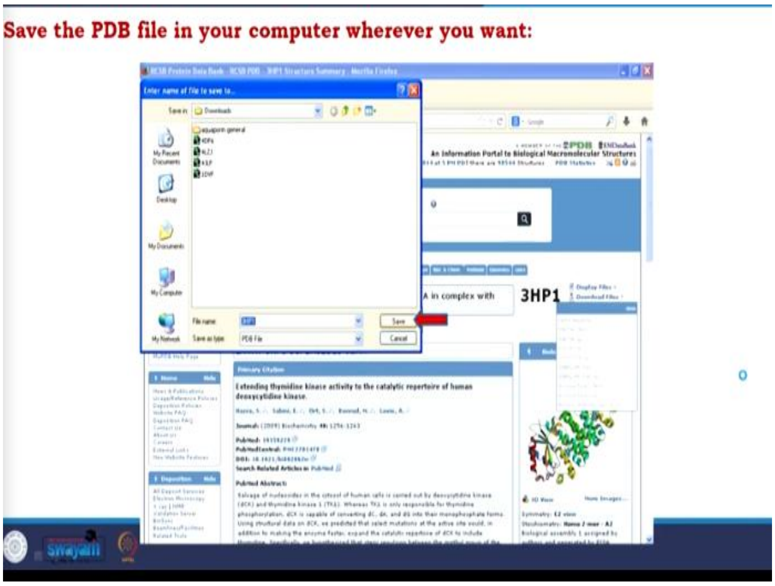
[OK] [Cancel]

Crystal structure of human dCK R104M/D133A in complex with L-dT and ADP

ENTRY 3HP1 SUPERSEDES 3EKK

3HP1

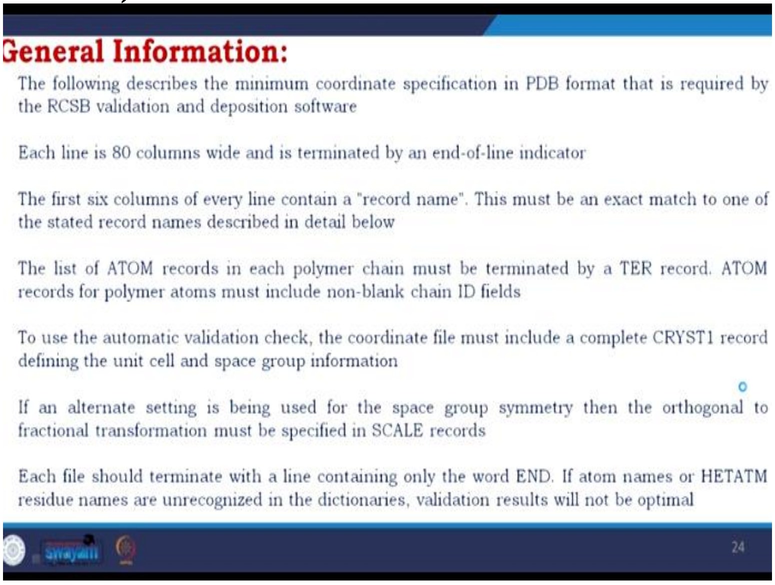
**(Refer Slide Time: 32:43)**



Save the PDB file in your computer wherever you want:

So, now we get the PDB 3HP1 PDB N, we save it to files, so you could do the PDB format or in the format where that data because it is big data, retailed reduced in zipped or anything.

**(Refer Slide Time: 33:07)**



General Information:

The following describes the minimum coordinate specification in PDB format that is required by the RCSB validation and deposition software

Each line is 80 columns wide and is terminated by an end-of-line indicator

The first six columns of every line contain a "record name". This must be an exact match to one of the stated record names described in detail below

The list of ATOM records in each polymer chain must be terminated by a TER record. ATOM records for polymer atoms must include non-blank chain ID fields

To use the automatic validation check, the coordinate file must include a complete CRYST1 record defining the unit cell and space group information

If an alternate setting is being used for the space group symmetry then the orthogonal to fractional transformation must be specified in SCALE records

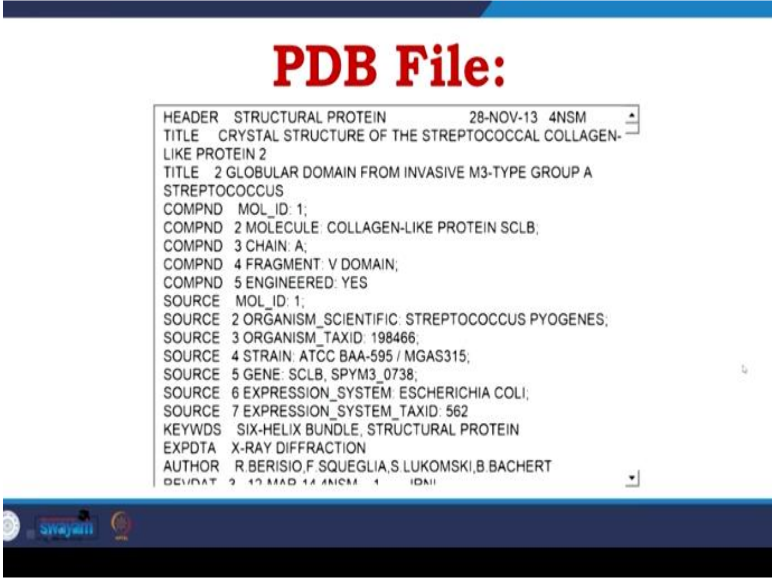
Each file should terminate with a line containing only the word END. If atom names or HETATM residue names are unrecognized in the dictionaries, validation results will not be optimal

The general information the following describes the minimum coordinate specification and PDB format that is required by the RCSB validation and deposition. What are they each line is 80 columns wide and is terminated by an end-of-line indicator. The first 6 columns of every line contain the record name, this must be an exact match to one of the stated record names describing the detail below. So, in detailing whatever you are doing, you have to record them.

The list of ATOM records in each polymer chain must be terminated by a TER record. Atom records for polymer atoms must include non-blank chain ID fields. To use the automatic validation check the coordinate file must include a complete list of one record defining the unit cell and the space group information. If an alternate setting is being used for the space group symmetry, then the orthogonal to fractional transformation must be specified in scale records.

Each file should terminate with a line containing only the word end. If atom names up HETATM residue names are unrecognized in the dictionaries, validation results will not be optimal. So, what are they? You had the rules but now I will explain it going to the actual PDB file.

**(Refer Slide Time: 34:34)**

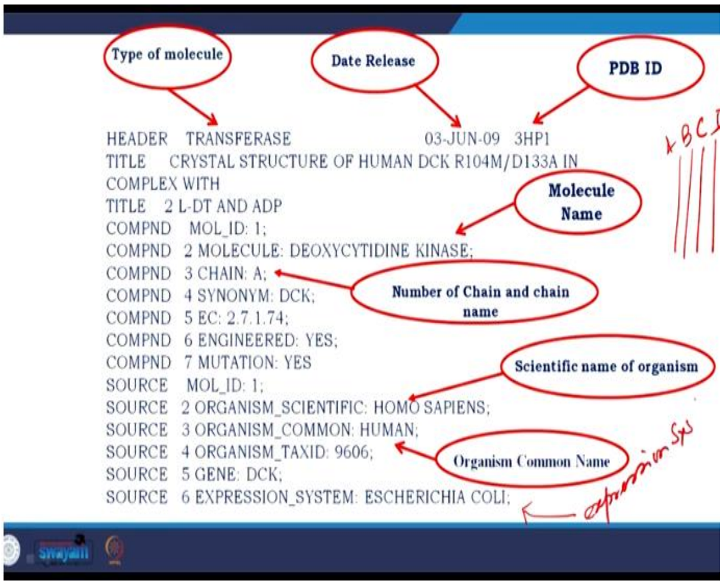


PDB File:

```
HEADER   STRUCTURAL PROTEIN                 28-NOV-13   4NSM
TITLE    CRYSTAL STRUCTURE OF THE STREPTOCOCCAL COLLAGEN-
LIKE PROTEIN 2
TITLE   2 GLOBULAR DOMAIN FROM INVASIVE M3-TYPE GROUP A
STREPTOCOCCUS
COMPND   MOL_ID: 1;
COMPND  2 MOLECULE: COLLAGEN-LIKE PROTEIN SCLB;
COMPND  3 CHAIN: A;
COMPND  4 FRAGMENT: V DOMAIN;
COMPND  5 ENGINEERED: YES
SOURCE   MOL_ID: 1;
SOURCE  2 ORGANISM_SCIENTIFIC: STREPTOCOCCUS PYOGENES;
SOURCE  3 ORGANISM_TAXID: 198466;
SOURCE  4 STRAIN: ATCC BAA-595 / MGAS315;
SOURCE  5 GENE: SCLB, SPYM3_0738;
SOURCE  6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE  7 EXPRESSION_SYSTEM_TAXID: 562
KEYWDS   SIX-HELIX BUNDLE, STRUCTURAL PROTEIN
EXPDTA   X-RAY DIFFRACTION
AUTHOR   R.BERISIO,F.SQUEGLIA,S.LUKOMSKI,B.BACHERT
REVDAT  2  12-MAR-14 4NSM    1       JRNL
```

So, this is a PDB file, and the reason I have given it if you click it here, you will see that it is a huge file. I hope you could understand it is a huge file how big it is? For an average PDB file, if you copy the file and paste it to your Word or notepad and make the font 10 to 11, you will get 50 to 60 pages. That is one of the reasons a lot of people do not read the PDB file. Whereas PDB file not only most people get the coordinate information important, but PDB file provides you much more than that, which we are going to take a look at.

**(Refer Slide Time: 35:49)**



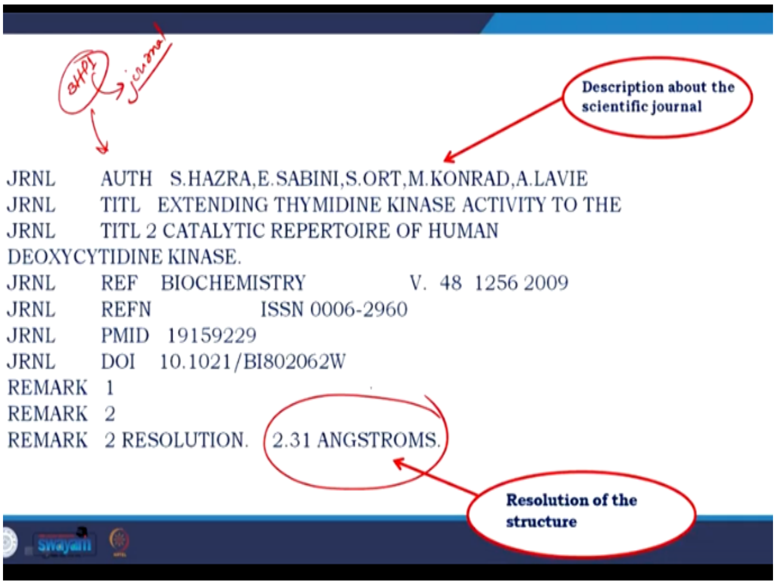
Type of molecule
Date Release
PDB ID
Molecule Name
Number of Chain and chain name
Scientific name of organism
Organism Common Name
+BCD
expression Sys

```
HEADER   TRANSFERASE                    03-JUN-09   3HP1
TITLE    CRYSTAL STRUCTURE OF HUMAN DCK R104M/D133A IN
COMPLEX WITH
TITLE   2 L-DT AND ADP
COMPND   MOL_ID: 1;
COMPND  2 MOLECULE: DEOXYCYTIDINE KINASE;
COMPND  3 CHAIN: A;
COMPND  4 SYNONYM: DCK;
COMPND  5 EC: 2.7.1.74;
COMPND  6 ENGINEERED: YES;
COMPND  7 MUTATION: YES
SOURCE   MOL_ID: 1;
SOURCE  2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE  3 ORGANISM_COMMON: HUMAN;
SOURCE  4 ORGANISM_TAXID: 9606;
SOURCE  5 GENE: DCK;
SOURCE  6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
```

So, this is the start of the PDB file and you see so many things one, there are columns, which talk about what is this, and the header talks about the enzyme, if you see the type of molecule here it is a transfer is so deoxycytidine kinase, the transfer is the date of release. So, you deposit the PDB information, and they give you the validation report, but you could hold it. So, you could hold it for the publication to be accepted so, finally, it was released on 03 Jun 09.

The PDB ID I talked about the title talk about the title of a crystal structure, then the chain, which is the number of chains and chain name this will tell how many polymers are there for polymer chains A, B, C, and D like that scientific name of the organism from where the protein is coming deoxycytidine kinase the human protein. So, it is Homo sapiens it is got the common name is human weight expressed in the expression system. So, a lot of experimental information is there expression system is E.coli so, though it is a eukaryotic protein, it is expressed in the prokaryote the standard system we use is E.coli.

**(Refer Slide Time: 37:31)**



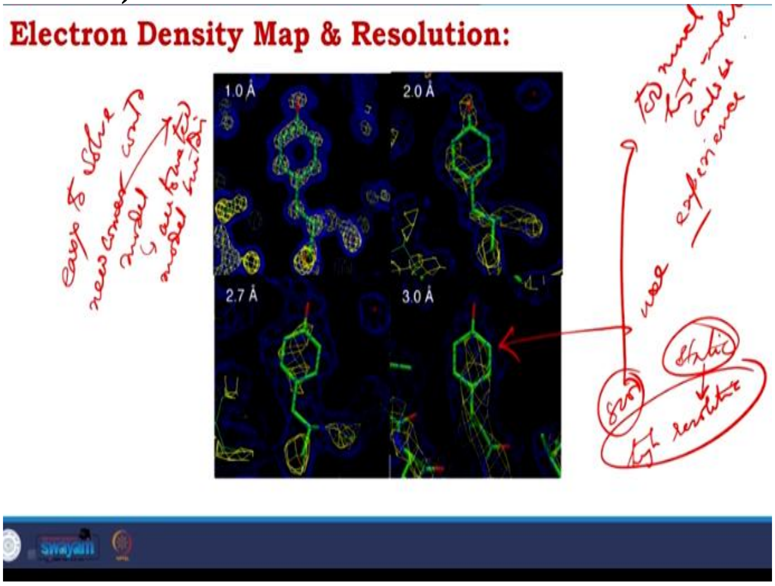
Description about the scientific journal

```
JRNL       AUTH   S.HAZRA,E.SABINI,S.ORT,M.KONRAD,A.LAVIE
JRNL       TITL   EXTENDING THYMIDINE KINASE ACTIVITY TO THE
JRNL       TITL 2 CATALYTIC REPERTOIRE OF HUMAN
DEOXYCYTIDINE KINASE.
JRNL       REF    BIOCHEMISTRY              V.  48  1256 2009
JRNL       REFN                 ISSN 0006-2960
JRNL       PMID   19159229
JRNL       DOI    10.1021/BI802062W
REMARK   1
REMARK   2
REMARK   2 RESOLUTION.   2.31 ANGSTROMS.
```

Resolution of the structure

All the information you could see in the journal you will see the author's title, why journal information is important because you know about a structure 3PH1, but what is the importance of the structure? What is that structure talking about you will get that information from the journal. So, you have to combine the information you get from the PDB ID about the structure information and go to the journal where the author writes about the findings, why this structure is important, and how this structure is used for further exploration of science.

**(Refer Slide Time: 38:23)**



Electron Density Map & Resolution:

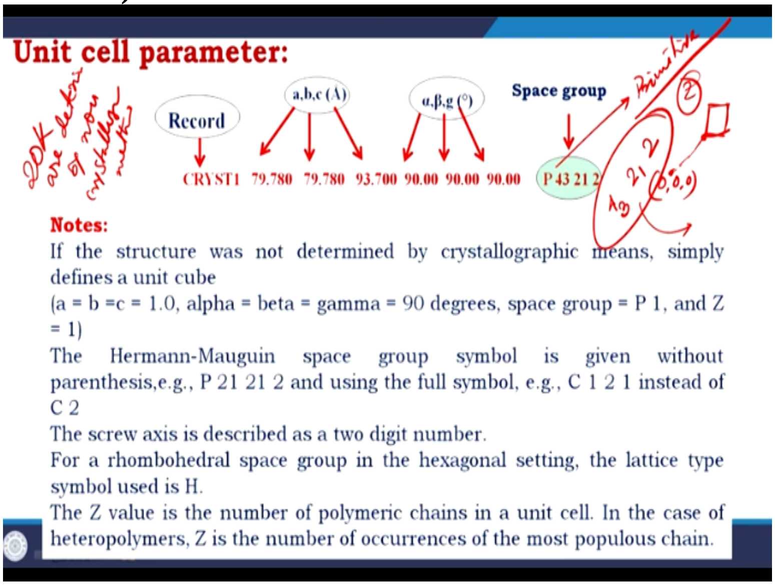
The resolution, resolution is very important why I talked about that, but just to a quick refresh, if you see, this is the same structure, but we get it in different 1 angstrom, 2 angstroms 2.7 angstroms, 3 angstroms and you see when you get in one angstrom It is so beautiful, it is so confident that it is easy to solve and even a newcomer could, could model it using automated modern building whereas when you go here you have to use experience.

So, the high-resolution structure is always easy, especially when you are going to look at the interaction between an enzyme and a substrate, or atomic-level interactions are required. I do not want to complicate here, but sometimes a very high-resolution structure below the number 1.5 Angstrom could be problematic because it comes with too much information. We all say that, when you look at a crystal structure, it gives you a static representation.

But now in the course, I have talked about you know understanding that a crystal having a lot of channels the small molecule could change conformation, so, it is not exactly static, here the protein loops are moving and also there is considerable flexibility. But when you are modeling you prefer one high-resolution confirmation with more high resolution, you might get too much information that you do not require. So, high resolution is always good but too much high resolution could be a problem.

If you download structures, especially with 1 angstrom you will see a lot of alternative confirmations, when you go and see the electron density map how we will look I will talk about these in the next class.

**(Refer Slide Time: 41:02)**



## Unit cell parameter:

Record
a,b,c (Å)
α,β,g (°)
Space group
CRYST1 79.780 79.780 93.700 90.00 90.00 90.00
P 43 21 2
20K deter are of non crystallgr meth
Primitive
2
$4_3$
$2_1$
2
(0,0,0)

**Notes:**

If the structure was not determined by crystallographic means, simply defines a unit cube

(a = b =c = 1.0, alpha = beta = gamma = 90 degrees, space group = P 1, and Z = 1)

The Hermann-Mauguin space group symbol is given without parenthesis,e.g., P 21 21 2 and using the full symbol, e.g., C 1 2 1 instead of C 2

The screw axis is described as a two digit number.

For a rhombohedral space group in the hexagonal setting, the lattice type symbol used is H.

The Z value is the number of polymeric chains in a unit cell. In the case of heteropolymers, Z is the number of occurrences of the most populous chain.

So, coming to that unit cell parameter, which is called the CRYST. In the CRYST information you will see the record portion it says Chris one you will get a, b, c you will get alpha, beta, gamma and you get space group. You need to understand something if the structure was not determined by crystallographic means, what will happen if you understand the problem? So, you have seen that among 180,000 or 75,000 structures, nearly 150,000 structures of the protein are solved by X-ray crystallography.

When it is the metal of extra crystallography, you need these because you have to connect the closest point of the unit cell with the start of the coordinates. And that is why you will need the crystal that a, b, c, and alpha, beta, and gamma and you have to develop a matrix to do that. But what if we know that around 20,000 structures are determined by a non-crystallographic method which NMR or cryo-electron microscopy we have talked about them?

So, what about them do they have creases to maintain the format they also have CRYST but it simply defines a unit cube where a b c = 1 and alpha beta gamma = 90 degrees. The space group P1 and Z = 1, the Hermann Maugin space group symbol is given without parentheses P 21 21 2 and using the full symbol C 1 2 1 instead of C 2. So, what are they? So, here if you see P, P stands for primitive remember we talked about all those symmetry things in crystallography.

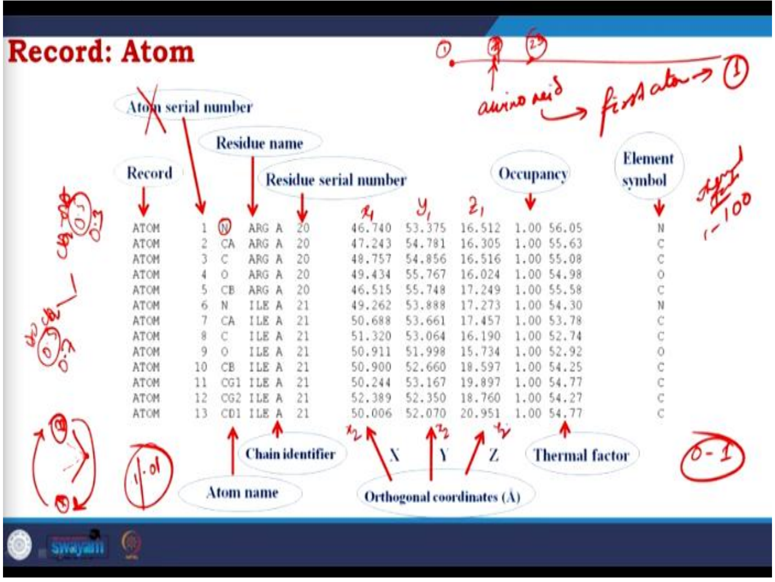
So, P is primitive 43 means, like these 21 if I right now, you will understand. So, it is the rotation and translation in different axis for full rotation followed by 3 / 4 cancellations, 2-fold rotation followed by half cancellation this way. The screw axis is described as a 2-digit number. For a rhombohedral space group in the hexagonal setting, the lattice type symbol used H. The Z value is the number of polymeric chains in a unit cell.

So, there are Z values also which is talked about the polymeric chains in a unit cell in the case of heteropolymers Z is the number of occurrences of the most populous chain. So, here we have one chain one polymeric chain all of you. Now, I want to talk about a very interesting thing you have learned about fee structure determination technique X-ray, NMR, and electron microscopy. Now, you know that you have around 150,000 structures determined by X-ray other 20,000 by NMR and cryo.

So, if I give you a PDB file, can you tell me the structure if it is not written there, how does the structure is solved? Can you tell me how the structure is solved? Is it solved by X-ray or electron microscopy or NMR? The first clue would be given from the CRYST if the creased is not human and the values are not exactly a = b = c = 1 and alpha beta gamma = 90. Then you know that it is solved by X-Ray crystallography majority of that problem is solved.

Now, how to differentiate between NMR and cryo? If you remember, I talked about the fact that in NMR, we do not determine the structure we average the structure. So, if you look at the NMR file, you will see multiple models there are generally 20 models represent the first one being the highest resolution. So, when you see no more chains present, you will be sure that they are NMR. So, this is a very simple trick to identify between the PDB files of determine by 3 techniques.

**(Refer Slide Time: 46:35)**



Record: Atom

```
ATOM      1  N   ARG A  20      46.740  53.375  16.512  1.00 56.05           N
ATOM      2  CA  ARG A  20      47.243  54.781  16.305  1.00 55.63           C
ATOM      3  C   ARG A  20      48.757  54.856  16.516  1.00 55.08           C
ATOM      4  O   ARG A  20      49.434  55.767  16.024  1.00 54.98           O
ATOM      5  CB  ARG A  20      46.515  55.748  17.249  1.00 55.58           C
ATOM      6  N   ILE A  21      49.262  53.888  17.273  1.00 54.30           N
ATOM      7  CA  ILE A  21      50.688  53.661  17.457  1.00 53.78           C
ATOM      8  C   ILE A  21      51.320  53.064  16.190  1.00 52.74           C
ATOM      9  O   ILE A  21      50.911  51.998  15.734  1.00 52.92           O
ATOM     10  CB  ILE A  21      50.900  52.660  18.597  1.00 54.25           C
ATOM     11  CG1 ILE A  21      50.244  53.167  19.897  1.00 54.77           C
ATOM     12  CG2 ILE A  21      52.389  52.350  18.760  1.00 54.27           C
ATOM     13  CD1 ILE A  21      50.006  52.070  20.951  1.00 54.77           C
```

Now, coming to the coordinates, so, as I told the atoms could be defined by recording the atom serial number is there. So, these are atoms’ serial numbers, but these are not residue serial numbers what is that remember? So, when we talk about atom serial number we are talking about the first atom it might not be so, if this is the protein it could start from here, it could start from here, it could start from here anywhere it starts the first amino acid, its first atom serial number 1.

But, when you come to residue serial number for these it is one for this these 2 like 7 for this it is 23 in that way. So, if you want to know whether the protein is fully crystallized or not, you cannot check the atomic serial number you have to check the residue serial number. These are the coordinates these are chain identifiers this is the atom name, this is occupancy and this is the thermal factor.

So, first, talk about these coordinates, this is where your atomic structure is, we will talk about this in detail, but occupancy and thermal factors are also very important information providers. What is occupancy? Occupancy is mostly you could explain it in X-ray why if you remember x-rays diffraction. So, if you are atom is here and a diffract so, when you calculate the electron density you get electron density in 1 point, so, the probability of finding the electron density is 1.

So, occupancy is the probability of finding the electron density of the atom which is ranging from 0 to 1 so, at the lowest it could be 0 at maximum it could be 1. Why it is very interesting as I told you so, if your protein amino acid has altered conformation suppose you have a serine so, CH 2 H 8 that serine has 2 conformations equally populated, the occupancy would be 0.5, 0.5 if one is more than 0.7, 0.3.

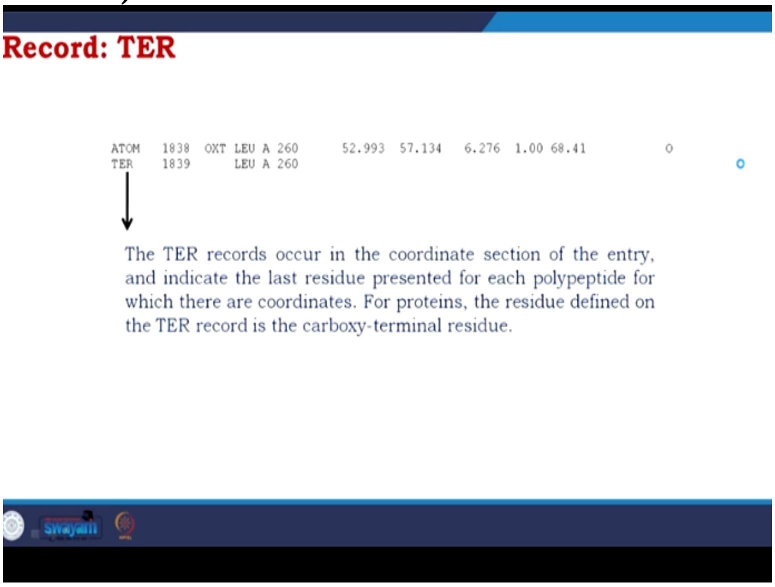
Now, here is a very interesting thing, and because, we are talking about the improvement of science, a lot of improvement could happen in this area, which will give us an idea about dynamics what is that I am talking about only crystallography here. See, if you think you have 1 amino acid here and 1 amino acid here like conformation, so, it adopts 1 confirmation here and another 1 here, but how it goes from one to another?

What I mean is if your protein one amino acid has confirmation A here and confirmation B here it should go from A to B or it should come from B to A. So, you should have some intermediate confirmations also and that is perfectly true. So, why it is not there, is because the software measuring has a cutoff and it will not provide you the measure where the contribution is 0.1, 0.01, or something there is the lowest cut-off.

But if you allow yourself to develop better software, you could identify other confirmations, which would be invaluable in understanding the dynamic nature of that residue you understand what I mean. So, you could not only identify confirmation you could also identify the transitions and some other confirmations which take an important part while this transition is happening. So, that is the importance of occupancy.

What is thermal factor thermal factor gives you an idea about the disturbance present there what I mean is if there is a loop that is very flexible and is moving randomly you are supposed to get a high thermal factor, ranging from 1 to 100.

**(Refer Slide Time: 52:07)**



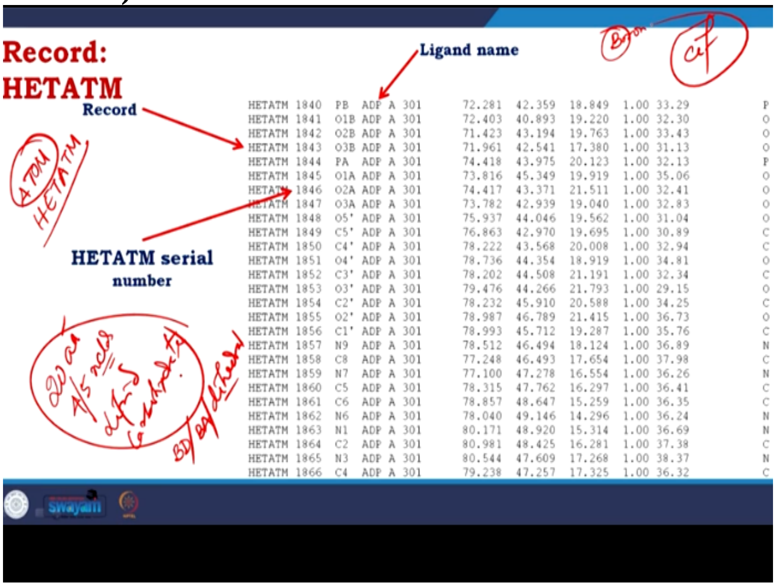
## Record: TER

```
ATOM   1838  OXT LEU A 260      52.993  57.134   6.276  1.00 68.41           O
TER    1839      LEU A 260
```

The TER records occur in the coordinate section of the entry, and indicate the last residue presented for each polypeptide for which there are coordinates. For proteins, the residue defined on the TER record is the carboxy-terminal residue.

As I talked about, there is a TER, the TER records occur in the coordinate safe section of the entry and indicate the last residue present and for each polypeptide for which there are coordinates. For proteins the residue defined on the TER record is a carboxy-terminal residue because it is N 2 C always.

**(Refer Slide Time: 52:28)**



## Record: HETATM

Record

HETATM serial number

Ligand name

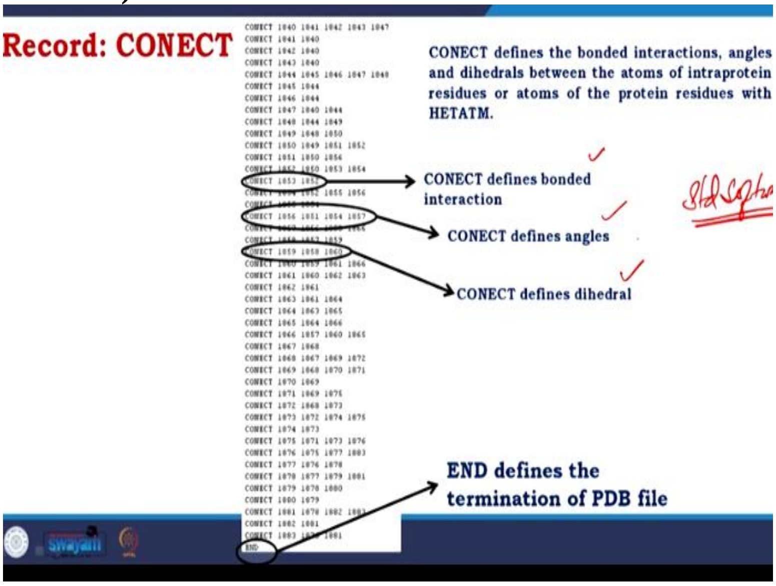
```
HETATM 1840  PB  ADP A 301      72.281  42.359  18.849  1.00 33.29           P
HETATM 1841  O1B ADP A 301      72.403  40.893  19.220  1.00 32.30           O
HETATM 1842  O2B ADP A 301      71.423  43.194  19.763  1.00 33.43           O
HETATM 1843  O3B ADP A 301      71.961  42.541  17.380  1.00 31.13           O
HETATM 1844  PA  ADP A 301      74.418  43.975  20.123  1.00 32.13           P
HETATM 1845  O1A ADP A 301      73.816  45.349  19.919  1.00 35.06           O
HETATM 1846  O2A ADP A 301      74.417  43.371  21.511  1.00 32.41           O
HETATM 1847  O3A ADP A 301      73.782  42.939  19.040  1.00 32.83           O
HETATM 1848  O5' ADP A 301      75.937  44.046  19.562  1.00 31.04           O
HETATM 1849  C5' ADP A 301      76.863  42.970  19.695  1.00 30.89           C
HETATM 1850  C4' ADP A 301      78.222  43.568  20.008  1.00 32.94           C
HETATM 1851  O4' ADP A 301      78.736  44.354  18.919  1.00 34.81           O
HETATM 1852  C3' ADP A 301      78.202  44.508  21.191  1.00 32.34           C
HETATM 1853  O3' ADP A 301      79.476  44.266  21.793  1.00 29.15           O
HETATM 1854  C2' ADP A 301      78.232  45.910  20.588  1.00 34.25           C
HETATM 1855  O2' ADP A 301      78.987  46.789  21.415  1.00 36.73           O
HETATM 1856  C1' ADP A 301      78.993  45.712  19.287  1.00 35.76           C
HETATM 1857  N9  ADP A 301      78.512  46.494  18.124  1.00 36.89           N
HETATM 1858  C8  ADP A 301      77.248  46.493  17.654  1.00 37.98           C
HETATM 1859  N7  ADP A 301      77.100  47.278  16.554  1.00 36.26           N
HETATM 1860  C5  ADP A 301      78.315  47.762  16.297  1.00 36.41           C
HETATM 1861  C6  ADP A 301      78.857  48.647  15.259  1.00 36.35           C
HETATM 1862  N6  ADP A 301      78.040  49.146  14.296  1.00 36.24           N
HETATM 1863  N1  ADP A 301      80.171  48.920  15.314  1.00 36.69           N
HETATM 1864  C2  ADP A 301      80.981  48.425  16.281  1.00 37.38           C
HETATM 1865  N3  ADP A 301      80.544  47.609  17.268  1.00 38.37           N
HETATM 1866  C4  ADP A 301      79.238  47.257  17.325  1.00 36.32           C
```

Then, hetero atom here also you get the record you get the hetero atom serial number, Ligand. First, tell me why we need ATOM and hetero atoms. So, when you get here you get the atom you get the hetero atom, what is the difference? Very very significant see when you are working with a protein molecule, the protein molecule of 20 amino acids. When you work with a nucleoside RNA, or DNA they actually 5 different nucleotides, there are liquid molecules there are carbohydrates, we have discussed them and you know all of them.

So, is it possible to know about their bond distance, bond angle dihedral, and all the information collected in a library? So, when you bring any protein, the protein’s geometrical character is already there in the software in the form of a standard library. Suppose, you are designing a new drug you include boron as a part of the molecule how the library would understand that if the library have to cover all your imagination?

Then that library should include all the chemistry which is impossible so, ATOM are the standard molecules that are here. hetero atom is designed by us or some unnatural molecules, where the geometrical parameters would be developed by you. And when you want to work with the software, you have to include that geometrical file in the form of a chemical information file or crystallography information file to the software you have to provide software that is called a hetero atom.

**(Refer Slide Time: 55:07)**



Record: CONECT

```
CONECT 1040 1041 1042 1043 1047
CONECT 1041 1040
CONECT 1042 1040
CONECT 1043 1040
CONECT 1044 1045 1046 1047 1048
CONECT 1045 1044
CONECT 1046 1044
CONECT 1047 1040 1044
CONECT 1048 1044 1049
CONECT 1049 1048 1050
CONECT 1050 1049 1051 1052
CONECT 1051 1050 1056
CONECT 1052 1050 1053 1054
CONECT 1053 1052
CONECT 1054 1052 1055 1056
CONECT 1055 1054
CONECT 1056 1051 1054 1057
CONECT 1057 1056 1058 1066
CONECT 1058 1057 1059
CONECT 1059 1058 1060
CONECT 1060 1059 1061 1066
CONECT 1061 1060 1062 1063
CONECT 1062 1061
CONECT 1063 1061 1064
CONECT 1064 1063 1065
CONECT 1065 1064 1066
CONECT 1066 1057 1060 1065
CONECT 1067 1068
CONECT 1068 1067 1069 1072
CONECT 1069 1068 1070 1071
CONECT 1070 1069
CONECT 1071 1069 1075
CONECT 1072 1068 1073
CONECT 1073 1072 1074 1075
CONECT 1074 1073
CONECT 1075 1071 1073 1076
CONECT 1076 1075 1077 1083
CONECT 1077 1076 1078
CONECT 1078 1077 1079 1081
CONECT 1079 1078 1080
CONECT 1080 1079
CONECT 1081 1078 1082 1083
CONECT 1082 1081
CONECT 1083 1076 1081
END
```

CONECT defines the bonded interactions, angles and dihedrals between the atoms of intraprotein residues or atoms of the protein residues with HETATM.

CONECT defines bonded interaction ✓

CONECT defines angles ✓

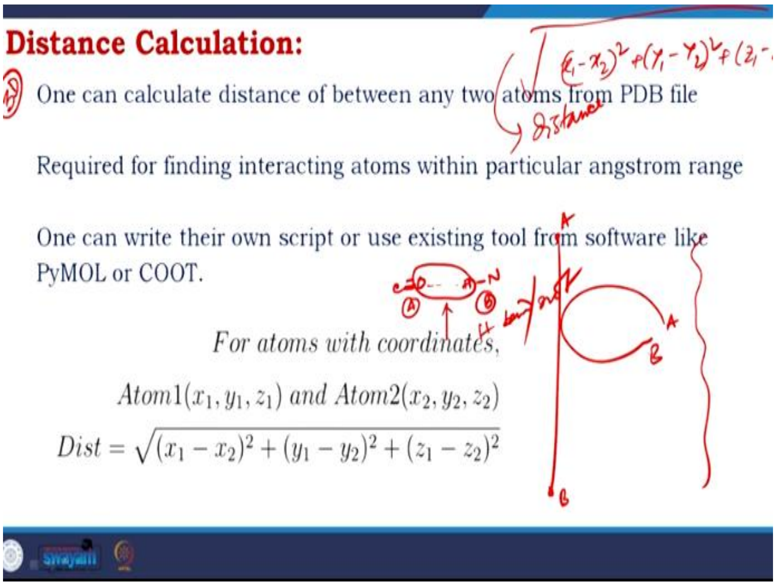
CONECT defines dihedral ✓

END defines the termination of PDB file

Then coming to the record called CONECT, CONECT defines the bonded interaction angles and dihedrals between the atoms of intraprotein residues or atoms of the protein residues with the hetero atom. So, if you see CONECT defines bonded interaction here, CONECT defines angles here, CONECT defines dihedral here, but when I am showing you, this is good for you to learn, you do not have to learn them why, because now all law standard software’s they already calculate this and take them immediately.

So, you do not need to habit the CONECT file because the CONECT file would be calculated and created by the software at the end of the PDB file you have to write END.

**(Refer Slide Time: 56:02)**



## Distance Calculation:

One can calculate distance of between any two atoms from PDB file

Required for finding interacting atoms within particular angstrom range

One can write their own script or use existing tool from software like PyMOL or COOT.

*For atoms with coordinates,*

$$Atom1(x_1, y_1, z_1) \text{ and } Atom2(x_2, y_2, z_2)$$

$$Dist = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

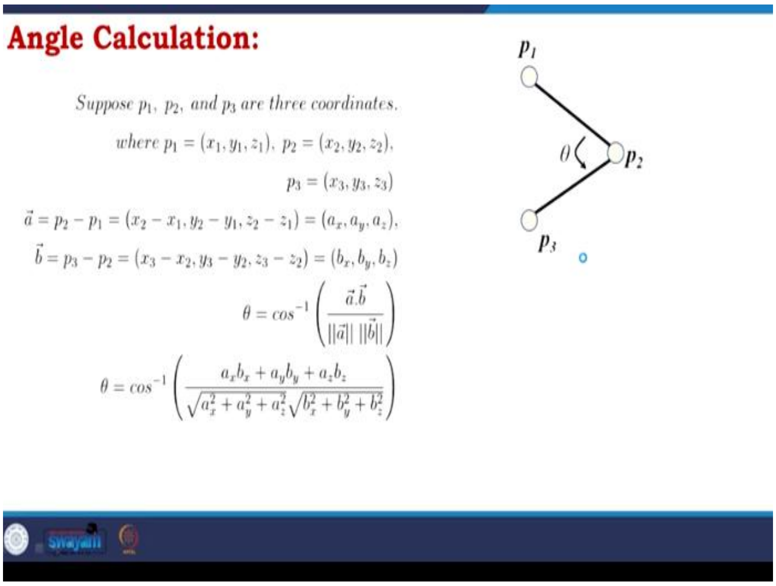
$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-}$
Distance
C=O
N
A
B
H bond?
A
B
A
B

So, what would be the potential use of this PDB file one distance calculation, one can calculate the distance between 2 atoms. So, if Atom1 has the coordinates x 1, y 1, z 1 and Atom2 have x 2, y 2, z 2 if you go back with it quickly you will see that here you have x 1 your coordinates y 1, z 1 this is x 2, y 2, z 2 then you get these values and you make this formula x1 – x2 whole square the distance y 1 – y 2 whole square + z 1 – z 2 whole square root and you get the distance.

So, what about the distance one can calculate the distance between any 2 atoms remember, this is 3D distance what I mean by that is if you have a peptide like this you could get the straight distance if the sequence is like that. So, let us say this is the length, but this is now this. So, the 2D distance is 2D the distance between A and B is this, but the 3D distance between A and B is this that is what is important here. It is required for finding interacting atoms within a particular Angstrom range.

One can write their script or use existing tools from software like PyMOL or COOT. So, what I am talking about, is if you are looking at a protein and you are looking at the 2 residues, that distance could be calculated by this, how it would be helpful when you see that you have an amino acids A and B if this one has C double bond O and this one have NH you could find the distance and talk about the fact that this is forming a hydrogen bond or not. In that way, many things could be attained by doing the distance calculation.

**(Refer Slide Time: 58:46)**



## Angle Calculation:

*Suppose* $p_1$, $p_2$, *and* $p_3$ *are three coordinates.*

*where* $p_1 = (x_1, y_1, z_1)$, $p_2 = (x_2, y_2, z_2)$,

$p_3 = (x_3, y_3, z_3)$

$$\vec{a} = p_2 - p_1 = (x_2 - x_1, y_2 - y_1, z_2 - z_1) = (a_x, a_y, a_z),$$

$$\vec{b} = p_3 - p_2 = (x_3 - x_2, y_3 - y_2, z_3 - z_2) = (b_x, b_y, b_z)$$

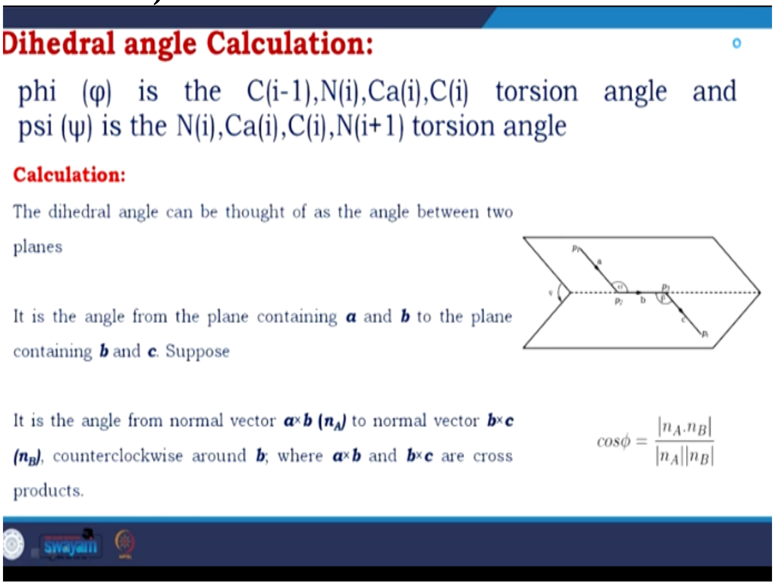
$$\theta = cos^{-1}\left(\frac{\vec{a}.\vec{b}}{||\vec{a}||\ ||\vec{b}||}\right)$$

$$\theta = cos^{-1}\left(\frac{a_x b_x + a_y b_y + a_z b_z}{\sqrt{a_x^2 + a_y^2 + a_z^2}\sqrt{b_x^2 + b_y^2 + b_z^2}}\right)$$

Similarly, Angle calculation whereas, distance calculation needs 2 points angle calculation would be having 3 points and you take 3 points p1, p2, and p3 which are having coordinates x 1, y 1, z 1 x 2, y 2, z 2 and x 3, y 3,z 3. Now, you make vectors a is p2 - p1 which is a x, a y, a z, b vector is p3 – p2 which is b x, b y,b z. So, now, the theta the angle would be cos inverse dot product of 2 vectors an absolute value of the 2 vectors.

So, why angle calculation is important to time you want to take a loop on how the active sides are aware, the molecules are interacting, how they are coming, how the movement happened and all, and the angle calculation always help in that.

**(Refer Slide Time 01:00:18)**



## Dihedral angle Calculation:

phi (φ) is the C(i-1),N(i),Ca(i),C(i) torsion angle and psi (ψ) is the N(i),Ca(i),C(i),N(i+1) torsion angle

**Calculation:**

The dihedral angle can be thought of as the angle between two planes

It is the angle from the plane containing ***a*** and ***b*** to the plane containing ***b*** and ***c***. Suppose

It is the angle from normal vector ***a***×***b*** (***n_A***) to normal vector ***b***×***c*** (***n_B***), counterclockwise around ***b***, where ***a***×***b*** and ***b***×***c*** are cross products.

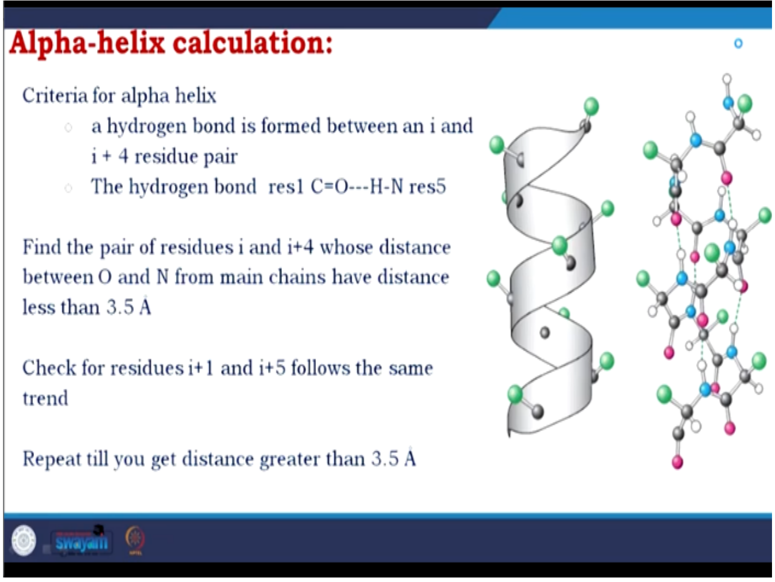
$$cos\phi = \frac{|n_A.n_B|}{|n_A||n_B|}$$

Dihedral angles There are many dihedral angles, but I will talk about phi and psi but also side changes are Kai's. So, anything you could calculate we have already talked about phi and psi phi is Ci – 1, Ni C alpha, and Ci. So, it is a torsion angle and psi is Ni, C alpha i, Ci, and Ni +

1. So, you could see the angles. So, the dihedral angle can be thought of as the angle between 2 planes we have already talked about when we are talking about protein.

So, I am quickly going through it is the angle from the planes containing a and b to the plane containing b and c if you see it here. It is the angle from normal vector a cross b to normal vector b cross c. In n A, n B is counterclockwise around b where a cross b and b cross c or cross products. So, by using those we could now calculate distance angle and dihedral angle and these provide you with the ability to calculate everything.

**(Refer Slide Time 01:01:48)**



## Alpha-helix calculation:

Criteria for alpha helix

- a hydrogen bond is formed between an i and i + 4 residue pair
- The hydrogen bond res1 C=O---H-N res5

Find the pair of residues i and i+4 whose distance between O and N from main chains have distance less than 3.5 Å

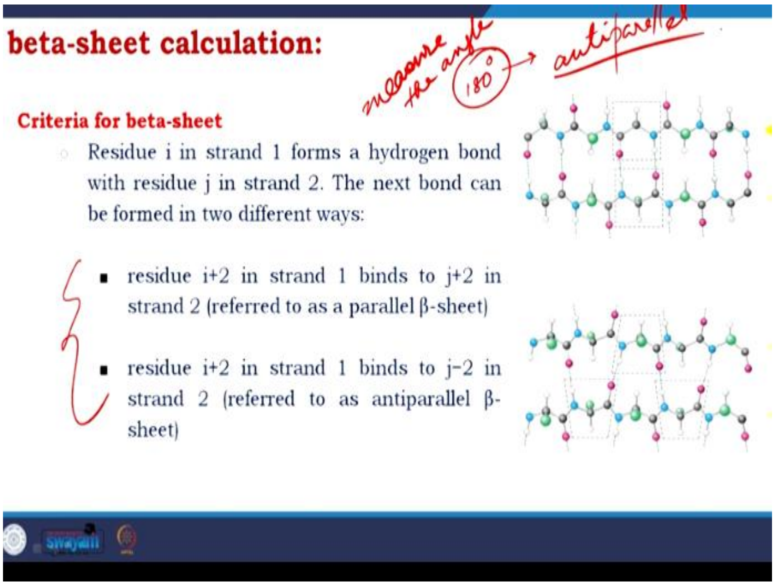
Check for residues i+1 and i+5 follows the same trend

Repeat till you get distance greater than 3.5 Å

And as I told you are the application you could calculate the existence of an alpha helix the criteria for an alpha helix a hydrogen bond is formed between i to i + 4 or n 2n + 4 and the hydrogen bond is between residue once carbonyl with residue fifth. I might find you have to find the pair of residues i and i + 4 whose distance between O and N from maintain distance less than 3.5 Armstrong.

Now, check for residues i + 1 and i + 5 that follow the same trend so, if it happens, then you know that the orientation you are going through is representing the alpha helix. Just think about understanding the distance calculation in 3D. You could have the ability to write an algorithm that will represent a program to calculate the presence of an alpha helix in a protein or not.

**(Refer Slide Time 01:02:54)**



beta-sheet calculation:

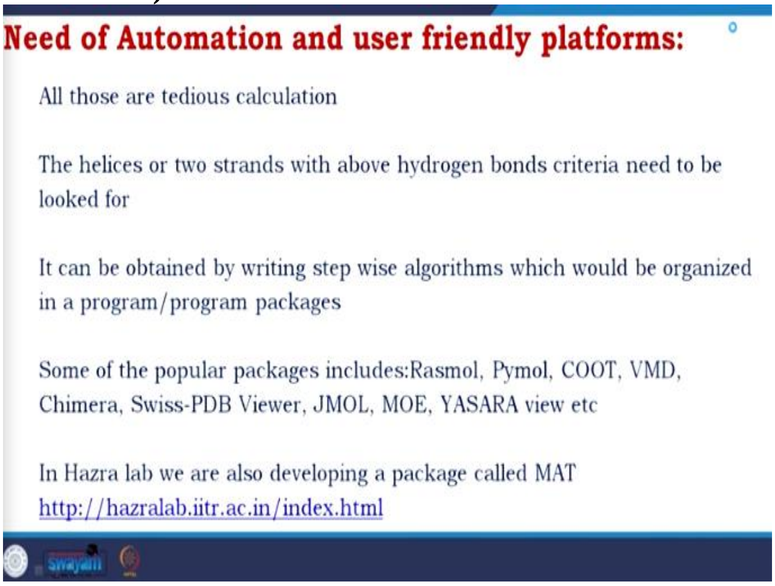
measure the angle 180° → antiparallel

Criteria for beta-sheet

- Residue i in strand 1 forms a hydrogen bond with residue j in strand 2. The next bond can be formed in two different ways:
  - residue i+2 in strand 1 binds to j+2 in strand 2 (referred to as a parallel β-sheet)
  - residue i+2 in strand 1 binds to j-2 in strand 2 (referred to as antiparallel β-sheet)

Beta sheet criteria for beta-sheet residue is in strand 1 forms the hydrogen-bonded residue j in strand 2. The next bond can be formed in 2 different ways residue i + 2 in strand 1 binds to j + 2 in strand 2, which is referred to as the parallel beta-sheet. Residue i + 2 in strand 1 binds to j - 2 in stand 2 refer to as an antiparallel beta-sheet. So, you could do that, in addition, you could measure the angle between i and j where the hydrogen bond is formed. If the angle is 180 degrees, it is antiparallel by incorporating those logics in the algorithm again you were going to have a program that could successfully calculate the presence of beta-sheet.

**(Refer Slide Time 01:04:02)**



Need of Automation and user friendly platforms:

All those are tedious calculation

The helices or two strands with above hydrogen bonds criteria need to be looked for

It can be obtained by writing step wise algorithms which would be organized in a program/program packages

Some of the popular packages includes:Rasmol, Pymol, COOT, VMD, Chimera, Swiss-PDB Viewer, JMOL, MOE, YASARA view etc

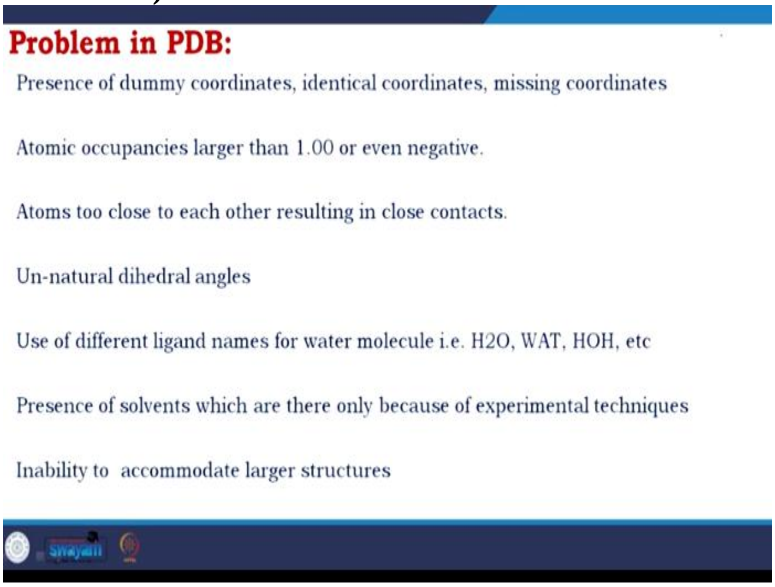
In Hazra lab we are also developing a package called MAT http://hazralab.iitr.ac.in/index.html

So, we could do that, but all those are very tedious calculations if you could try but please you should try to try their download the PDB file get the coordinates, and try those formulas. The helices of the 2 strands hydrogen bond criteria are the distances, distances between like and enzyme all these are very tedious calculations. So, it could be obtained by writing step wie algorithms you have to understand the logic and you use logic to make logical reasoning that logical reasoning could be converted to a program using any programming language.

So, I talked about some of the popular packages are RasMol, PyMOL, COOT, VMD, Chimera, Swiss-PDB viewer, JMOL, MOE, and YASARA view a lot of them are there because I am taking your class I would also love to introduce a software package currently developed by us and we are working with this my first Ph.D student guzzle c have already graduated working on.

So, what is our aim is we are developing a package that would be making the non - computational user coming from core biology giving a user-friendly platform I will talk more about this, but if you are interested, I have provided the link you could see that just one thing we have written this in C language and currently, this is the fastest parser present in the world the calculation the parsing, what we do is the performance wise is the fastest among the currently established software's. What is parsing I will talk a little bit about it.

**(Refer Slide Time 01:06:14)**



**Problem in PDB:**

Presence of dummy coordinates, identical coordinates, missing coordinates

Atomic occupancies larger than 1.00 or even negative.

Atoms too close to each other resulting in close contacts.

Un-natural dihedral angles

Use of different ligand names for water molecule i.e. H2O, WAT, HOH, etc

Presence of solvents which are there only because of experimental techniques

Inability to  accommodate larger structures

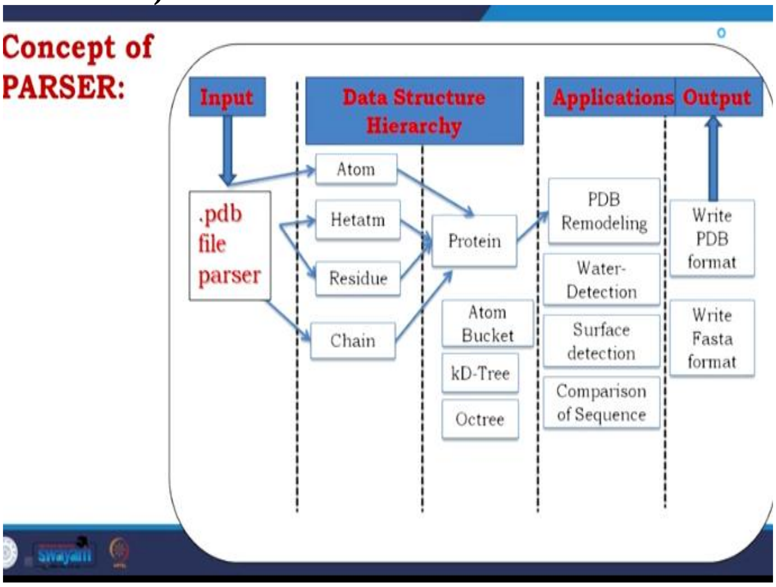
But before that what are the problems in PDB the presence of dummy coordinates, identical coordinates, and missing coordinates are the problem. Atomic occupancies larger than 1.00 or even negative which should not be there atoms too close to each other resulting in close contact breaking the rule of chemistry which should not be there, unnatural dihedral angles. Use of different ligands and names for water molecules, H2O, WAT, HOH, etcetera there should be one name.

It is easy for the program Presence of solvents which are there only because of experimental techniques mainly in crystallography to get crystals we use a lot of conditions and that brings many ligands to be incorporated or involved in the actual crystal structure. They are misleading most of the time they are creating a nonbiological environment inability to accommodate larger structures with the current advent of cryo-electron microscopy.

You have already seen huge structural assemblies coming the PDB format was created to accommodate crystal structures in 1970s. Very few or no one had imagined that time that cryo-electron could solve so big assemblies now. Even crystallographic methods have solved ribosome structures and cryo is reaching with mega Dalton level. So, we need to accommodate all those information coming from the largest structures and there are many of them.

So, how to solve them already PDB introduced mmCIF as I talked about, and they are also bringing mmCIF which our solution, but if some of you are interested to work with you need a very good architecture.

**(Refer Slide Time 01:08:20)**



Concept of PARSER:

Input

Data Structure Hierarchy

Applications

Output

.pdb file parser

Atom

Hetatm

Residue

Chain

Protein

Atom Bucket

kD-Tree

Octree

PDB Remodeling

Water-Detection

Surface detection

Comparison of Sequence

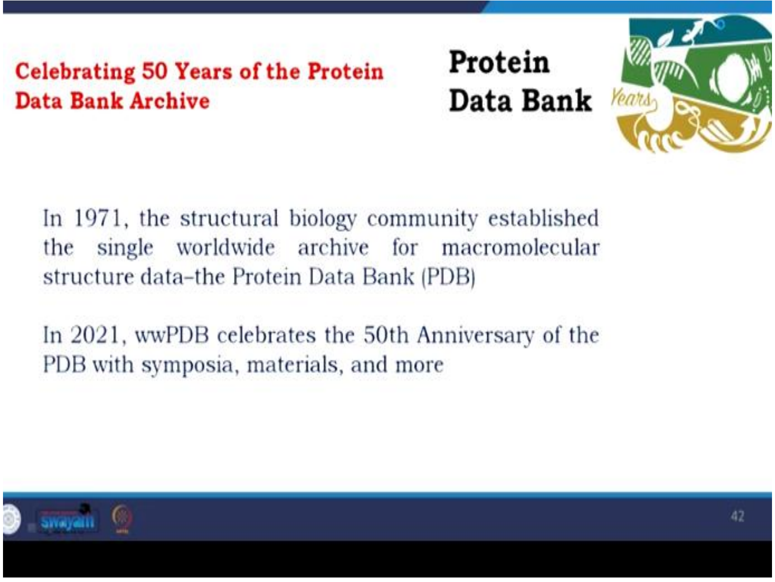
Write PDB format

Write Fasta format

So, you have input your data structure hierarchy, where you have the polymeric chain, then in the next level residue then in the next hetero atom then in the next level ATOM and you could have develop you know data structure hierarchy you could use existing data structures like atomic bucket, kD Tree, Octree and once you pick the protein and develop this hierarchy, you have the input of the PDB file the architecture when it performed that is called parsing.

So, you have to develop a parser this parser will take your protein will divide it into hierarchies and that would make it easier to provide further outputs right, like PDB remodeling, water detection, surface detection, the composition of sequence, and many more many 1000’s of application is possible. So, that is one where there are a lot of scopes of you know, future research because with enhancement with new techniques incorporating, we need to have some flexible architectures to come.

**(Refer Slide Time 01:09:43)**



Celebrating 50 Years of the Protein Data Bank Archive

Protein Data Bank

In 1971, the structural biology community established the single worldwide archive for macromolecular structure data–the Protein Data Bank (PDB)

In 2021, wwPDB celebrates the 50th Anniversary of the PDB with symposia, materials, and more

And with that, I would finish today's description today's class of PDB with the news that we are celebrating 50 years of the Protein Data Bank this year. And as you all know, now, we started in 1971. And in 2021, we are celebrating the 50 years of Protein Data Bank. So, if you are interested to go to the PDB search celebration of 50 years of Protein Data Bank you will see a lot of programs. And because of the corona situation, a lot of programs are becoming webinars.

So, you could have in or taken part in them, you could have gained further motivation through them. With that again, I finished this class, I again request you thank you for listening, I request you to continue listening to understand but at the end, do it on your own, go to the PDB file, go to RCSB download the PDB file, do all the calculations. Once you do it by your hand, you could have much more understanding. Thank you very much.