Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 16a
Protein Tertiary Structure I

In this lecture we will discuss about protein tertiary structures right. So, we already discussed about the protein primary structure. What is primary structure?

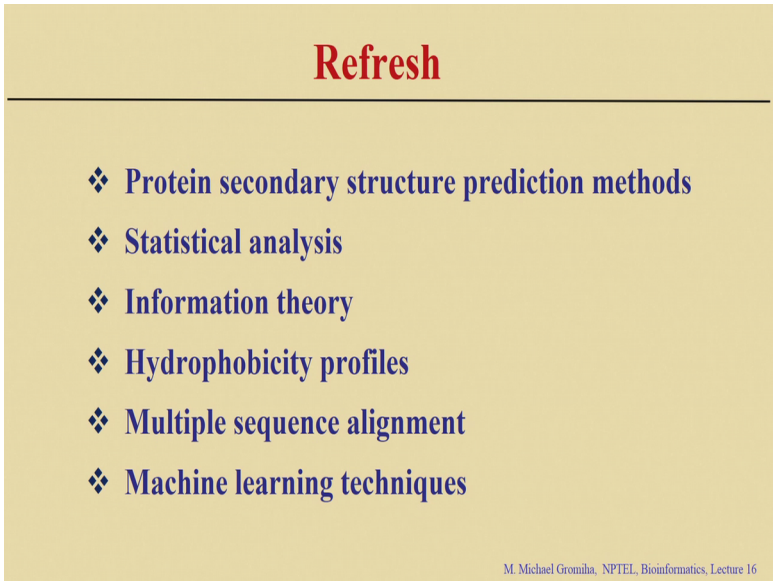
Student: Sequence.

Amino acid sequence. Then protein secondary structures, what are different secondary structures?

Student: Alpha helix beta strands

Alpha helix and beta strands right. So, then we derived several features or several properties from protein sequences as well as for the protein secondary structures. In the previous class we mainly focused on the different methods or predicting protein secondary structures. So, we discussed about statistical analysis.

(Refer Slide Time: 00:50)



For example, Chou and Fasman method, this based on the propensity of amino acid residues at any conformation either helix or strand right or coil. Then we discussed about

information theory is called GOR method right they use the information theory and the information at the central position as well as a window length of 17 residues left side 8 residues and right side 8 residues right and use the information right to predict the secondary structures. We discussed about hydrophobicity profiles, the plot connecting the sequence and the values right. So, either they get different patterns or you can take the average value and see the peaks to identify a secondary structure.

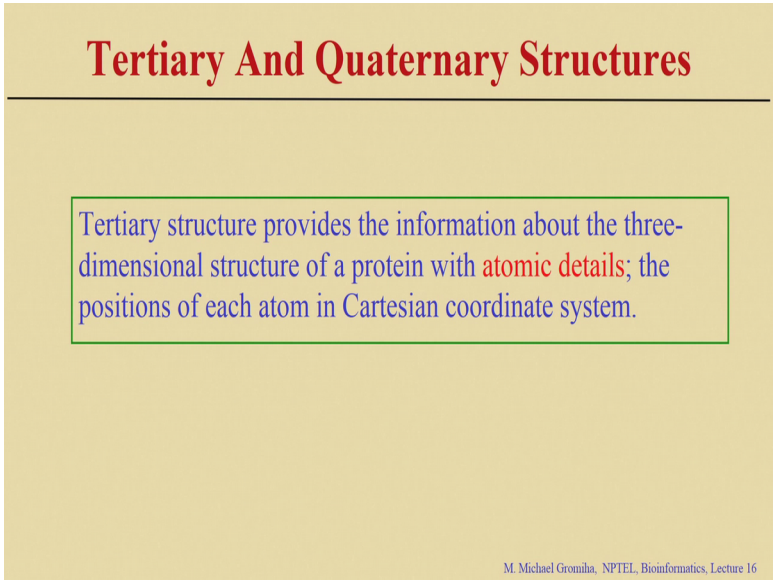
We discussed about multiple sequence alignment right this mainly based on the alignment of different sequences right when in form of either PSSM matrices or any other different methods right. So, use this alignment and predict the secondary structure either the use the GOR method or they can use the ensemble based information from known structures. Later on then we discuss about the machine learning techniques, they used all the information right and they train this information using machine learning techniques like neural networks or support vector machines or different methods right and you predict the secondary structure. We also discussed about the consensus method right the ensemble based method or the meta servers right either they take based on noting or train the output of different methods right to get the desired output.

So, this class we discuss in 3D structures right. So, what do you think about the 3D structures? Which information you can obtained from protein 3D structures?

Student: The orientation.

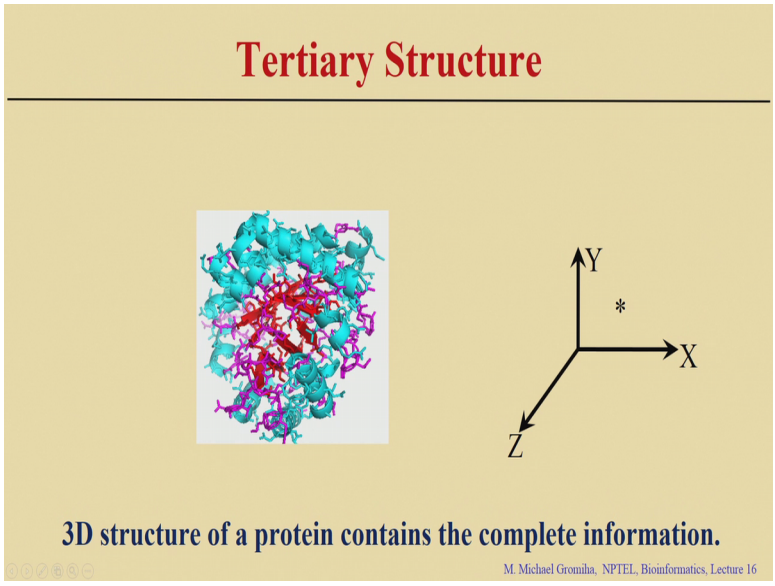
The orientation as well as specifically we consider coordinates.

(Refer Slide Time: 02:38)



So, tertiary structures are the 3D structures. So, provide the information about the three dimensional structure of a protein right with atomic details. So, it can give you the exact location of each atom right in a specific residues and a protein right X and Y Z coordinates in a Cartesian coordinate system. Then coordinate structures they provide the combination of the secondary tertiary structures having different sub units to give tertiary structures.

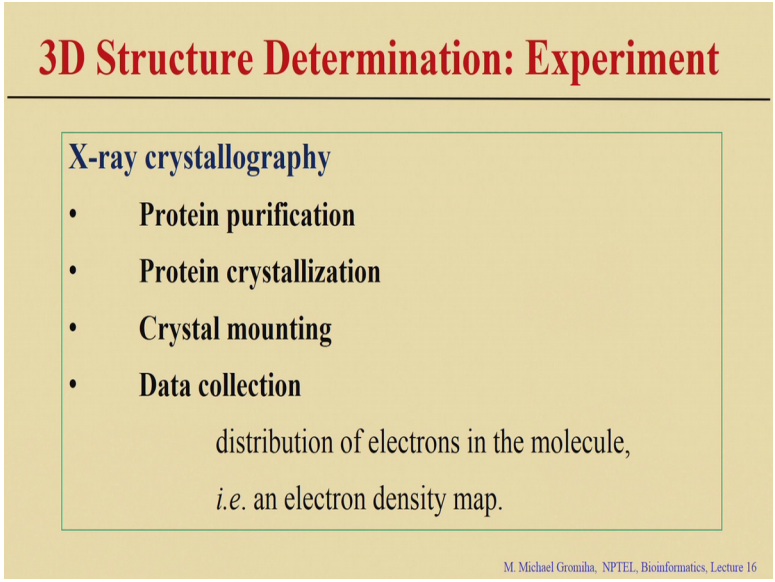
(Refer Slide Time: 03:03)



So, here if you see the protein, they can see the structure of a particular protein here. So, there are sequence some know are secondary structure helix you can see in the form of spirals and you can see some arrows here they represent the strands right. And then they interact with each other the atoms and the residues, the interact with each other based on the various types of interactions depending upon the residue types either they make the electrostatic interaction or van der Waals interaction or hydrophobic interactions and they form the unique 3D structures.

We talk about the three dimensional structure of a protein, it gives the atomic coordinates say if this is an atom here denoted here. So, give the information you got X, Y and Z coordinates. How to get this structures right to how to obtain the information regarding the 3D structures right, there is various experimental methods, like X-ray crystallography, NMR spectroscopy, electron microscopy, the different methods which help right to determine the 3D structures of proteins or other macromolecules even of small molecules.

(Refer Slide Time: 03:56)



So, if you look about the X-ray crystallography these one of the most prominent methods available in literature currently, we have the very sophisticated instruments right crystallographic instruments right. So, you can obtain the three dimensional structures right of even a big small molecule; big molecules if you are able to crystallize. In X-ray crystallography the principle used in X-ray crystallography is what.

Student: X-ray diffraction

X-ray diffraction right, if you have the X-ray diffraction right this is the basic principle used in X-ray crystallography. So, there various steps to use X-ray to get the structures you say X-ray crystallography. First we need to have a protein right. So, in that case either we have to synthesize the protein right or you can extract and then get the proteins, isolate the protein right. So, and if we get the protein then you have to ensure that this protein is pure. So, the various steps which are involved in the protein crystallography, first you need to get the pure protein. So, protein purification once we get the pure protein then we need to crystallize because the X-ray crystallography diffraction; it is important in order to get a diffraction we need a crystal. So, in this case we need to use the different experimental conditions so that you can make this protein in a crystallized form right. What is a crystal?

Student: Periodic arrangement of atoms.

The periodic arrangement of the atoms right in arrangement of atom there is with the same unit right for all this units are same right. So, in this case will get crystallized of a protein and then we need to next is crystal mounting is an also an important step currently in several laboratories like the synchrotron radiation and all it is automated. So, in this case it can automatically mount at different directions and you can get the diffraction pattern otherwise this is also very important step to mount the crystals in proper orientation and different proper directions.

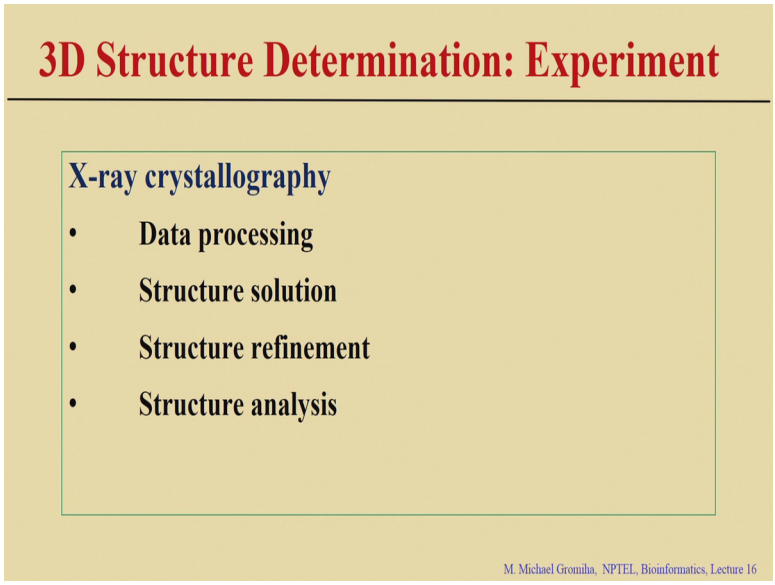
Once we do all this things then you can pass the X-rays right and you can get the diffraction pattern. So, if we have the crystal and you get the diffraction pattern this will give you, what is a meaning of the diffraction pattern?

Student: The position of the atoms.

Yeah, this is a distribution of the electrons molecules how they position in the molecules. So, they are diffracted right and then you can see the patterns. So, getting this patterns this patterns of the electrons right that is why we call as electron density map. So, you can you have a kind of maps right depending upon the population of the electrons the crystal right. So, you can get the electron density map. This is what we get using the

diffraction of this crystals. Then after that we need to process, electron density map right this is location of these electrons we need to process these data to get the structures.

(Refer Slide Time: 06:37)

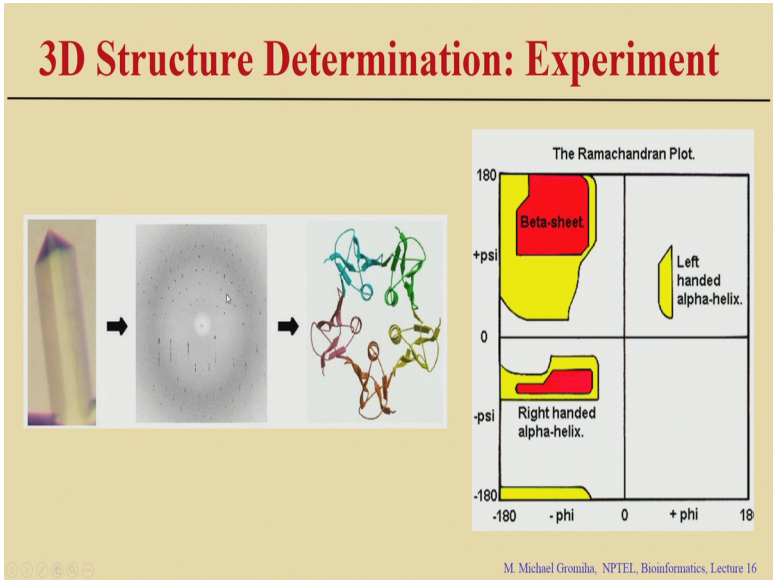


There various methods to process this data right, because we need to have the intensity or amplitude right as well as the phase.

So, there are this is the problem the X-ray crystallography. Now, there are different methods to solve the problem of this phase issue right and then you can obtain this structure. So, then once we get this process the data we will get the structures then we need to refined the structures because depending upon this quality of the structure right we need to match this two data right electron density map when their own structure, then they like to refine this structure and finally, we can get the refine structures.

Once we get the refine structures right then this structure can be useful for doing an analysis to understand the behavior of this protein structure, what is relationship between structure and function, where are the active residues, which residues are involved in this type of stability right and folding and so on, this is possible. So, this is X-ray crystallography; this is the major steps.

(Refer Slide Time: 07:32)



First we need the crystal right we have to get the crystal and get the diffraction you can give a pattern like this and you process this data right then finally, we get the 3D structures this what we need. Then we check this quality of this 3D structures right and validating this using various methods and one of the methods is the Ramachandran plot and see whether these residues are in the allowed regions mainly in the secondary structures alpha helixes and the a beta strands.

So, if you look into this known structures X-ray crystallography is one of the most prominent method right you can see a quite large number of structures in the Protein Data Bank I will show very soon and several researches they got Nobel Prize right for using X-ray crystallography to solve the structures. Starting from the beginning, the discovery of this X-rays and then using X-ray is to diffract on crystals right and then for getting this for the small molecules to get the structures like the Bragg and Bragg.

So, they showed that the path difference is the integral multiples of this wavelength. What is the famous equation for the Bragg's law?

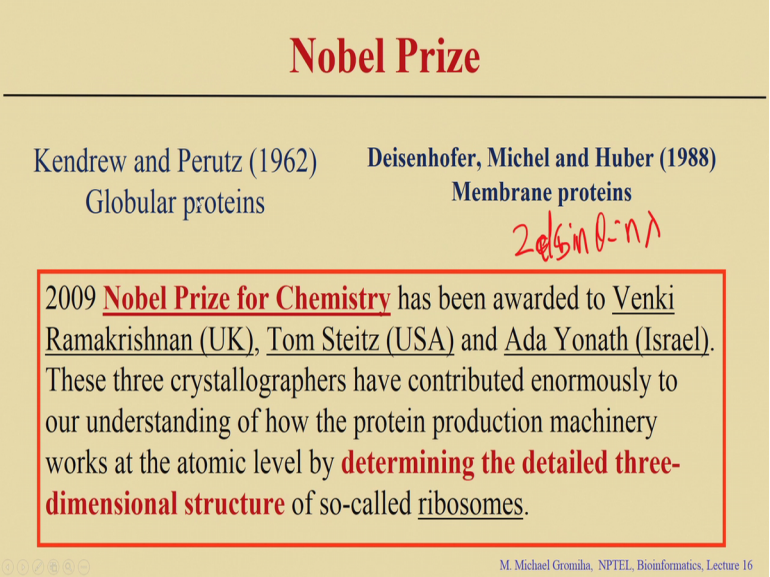
Student: $n\lambda$ equal to.

$2d \sin\theta$ equal to.

Student: $n\lambda$.

nλ, integral multiples of this wavelength this is the path difference or a θ is the angle between the incidence.

(Refer Slide Time: 08:42)



Student: And the.

And the plane right and what is d.

Student: The interplanar distance

This interplanar distance right these interplanar distance right, θ is the angle between the angle of incidence or under diffraction. So, λ is the wavelength. So, they showed that. So, 2d sinθ = n, nλ and we get the highest diffractions only if this integral multiples of lambda right. So, then they used this X-ray diffraction to solve the structures of proteins mainly for the globular proteins say myoglobin and the hemoglobin this was solved structures right, Kendrew and Perutz right they got Nobel Prize in 1962 for understanding the structures of globular proteins use the X-ray crystallography. Then there are several other research like the vitamin b12 complex and then they used for the DNA structures right. Who got the Nobel Prize in DNA structures?

Student: Watson and Crick

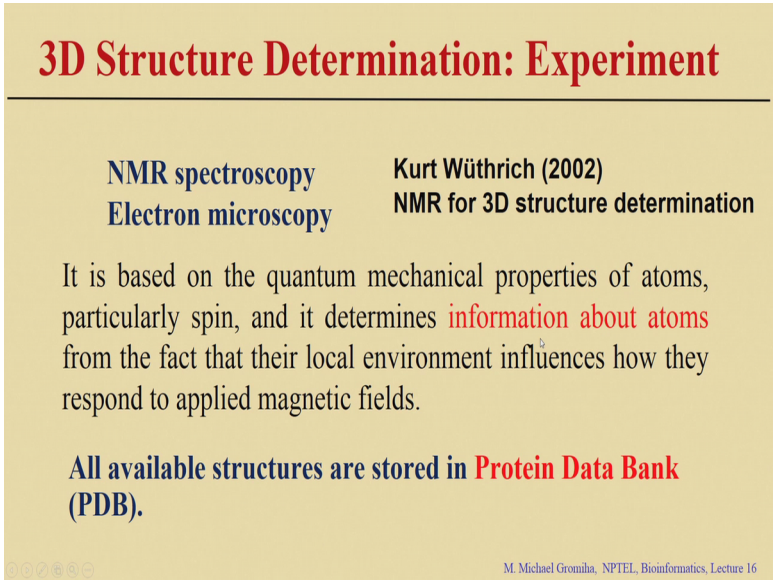
Watson and Crick got Nobel Prize and then they used protein-DNA interactions Klug got the Nobel Prize for the protein-DNA interactions. Then for the case of membrane

proteins right in 1988 Deisenhofer right Michel and Huber they solve the structures of the membrane; first membrane protein right the photosynthetic reaction center and they got Nobel Prize in 1988.

So, currently 2009, we got the Venki Ramakrishnan, Tom Seitz and Ada Yonath right they are three crystallographers right, they enormously contributed to understand the protein production machinery right at the atomic level using the 3D structures of ribosomes right. So, they got the Nobel Prize in 2009. Also there are Nobel Prize winners for the GPCRs right and the other works on these protein crystallography. So, X-ray crystallography is a major technique used to determine the 3D structures of protein molecules as well as other biological macromolecules.

Then we have also NMR spectroscopy and microscopy are used to determine the structures. If you look on the NMR spectroscopy, this based on the quantum mechanical properties of atoms like the spin that determines information of the atoms right and the local environment right with response to the magnetic field.

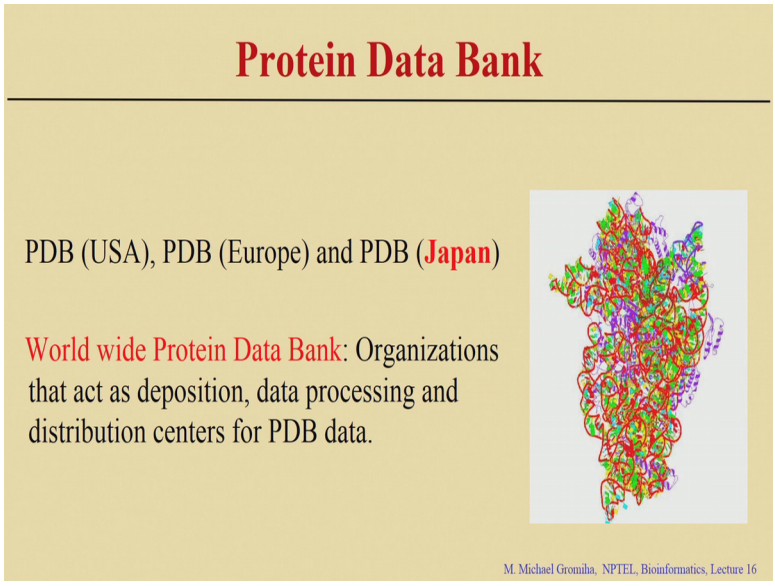
(Refer Slide Time: 10:45)



So, they have this spectrum and use the distance constraints in the NMR spectra and use distance constraints to develop models. This is why if you look into these structures for the crystallography they get one structure and NMR they create several models right you can see NMR structures they have 10 models and there is a average structure for among the all the different models they get the average structures.

So, in 2002 Kurt Wuthrich he got Nobel Prize for using NMR for 3D structure determination. So, the several techniques and the major one is X-ray crystallography NMR spectroscopy and currently electron microscopy is also used for determining the 3D structure. So, you solve all these structures and all the structures are now deposited in Protein Data Bank right PDB is called the Protein Data Bank right all the structures are now deposited in Protein Data Bank. In 1977, Brookhaven National Laboratory started to collect the structures of proteins; at that time there were only 15 structures right they gathered the structures and started to develop a data set Brookhaven National Laboratory. Later on with the structures coming up then they tried to collect the more structural again and again right and they formed the consortium, they form this Research Collaboratory in Structural Bioinformatics, they found this the Protein Data Bank.

(Refer Slide Time: 12:13)



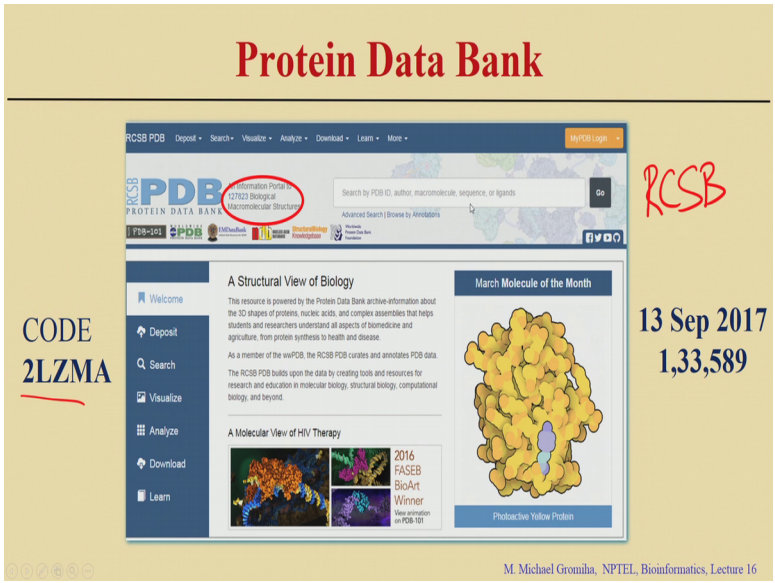
So, the Protein Data Bank like it is organized by the various institutions from different countries like Japan the Europe or the US. So, they have a PDB USA and PDB Europe as well as a PDB Japan. So, they share the data with each other and they collect the information and they maintain the structures of these protein three dimensions, 3D structures.

So, now, the growth of the PDB depends on the structures deposited in the Protein Data Bank right. Earlier days they accepted all these structures when they are published in the literature, if you see the latest ones and the previous ones latest one they have several

options and also they have to pass through several quality checks. And the older structures they did not perform the quality check there was there were several issues in the old entries. Secondly, in earlier days the crystallographers or the using NMR spectroscopy they solve the structures and they publish and then deposit in the databank. The current scenario first we have to get the code from the PDB before you publish your article. In this case the PDB right who is the maintain the PDB right they curating the PDB they check the quality of your structure and if you pass the quality check then only they assign the code.

So, the only after you getting the code you can publish you research article. So, in this way they are able to maintain the quality of the structures which are available in the Protein Data Bank. This is the reason why the PDB is the widely accepted the unique resource for getting the structures 3D structures of proteins and as well as other macromolecules and complexes. So, here I show the organization of the Protein Data Bank and the data which are available in the Protein Data Bank and how to extract the data and how to utilize the data from the Protein Data Bank.

(Refer Slide Time: 14:13)

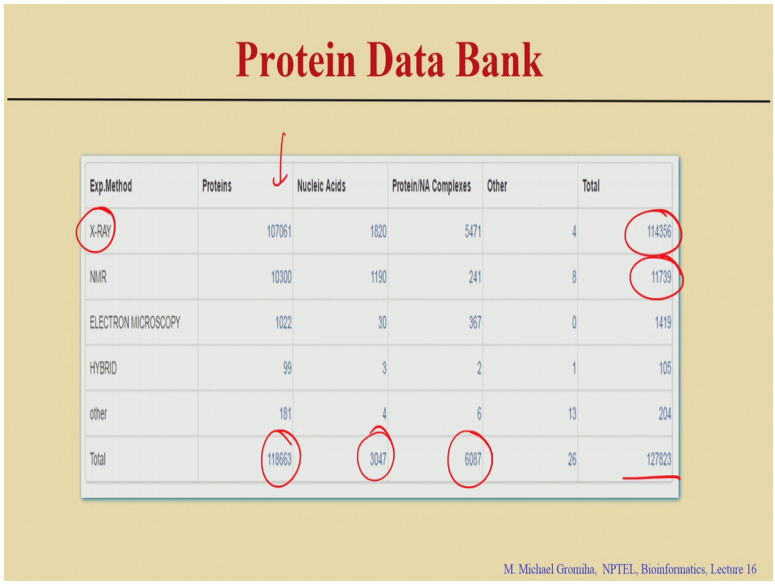


So, PDB is, PDB is available in the RCSB website right Research Collaboratory in Structural Bioinformatics. So, they maintain the website. So, it is condensed all these solved structures. So, currently it has about at the moment you have 133,589 structures and these structures is increasing like periodically right. So, every time you can see you

can see increase in the growth of the number of PDB structures available in the Protein Data Bank.

So, this is the front page of the Protein Data Bank right you can see the statistics right here and here you can see the search option. You can, use any search option right to get this here information from Protein Data Bank. So, you see here there are several options available to deposit or to search you can use advanced search right and you can visualize the molecule and you can analyze the molecule or you can download right. And there are several tutorials available in the Protein Data Bank which are related with the usage of protein 3D structures either there are several tools or you can several analysis or several aspects or we can do with these 3D structures of proteins.

(Refer Slide Time: 14:52)



So, here you can use the search option right you can if you want to have any specific protein you for example, this is the code for the lysozyme. So, 2LZM; this is the chain information if you give this one you will get this the data for this lysozyme.

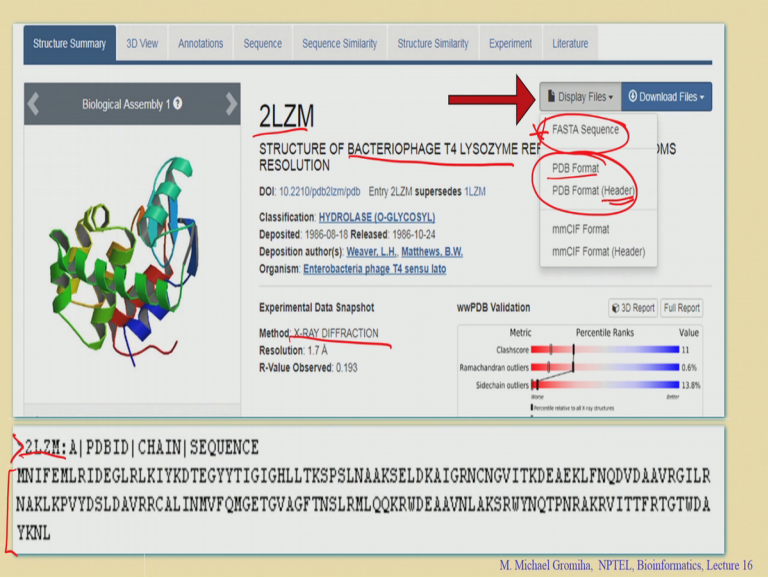
If you look into this statistics of Protein Data Bank, here you see as I discussed earlier X-ray crystallography is the most prominent technique to obtain the 3D structures of proteins. So, if you see among this 127,000 structures. So, you can see about 114,000 structures are obtained from X-ray crystallography. Then we see NMR spectroscopy there is about 11,000 structures right followed by the other techniques. And if you look into these structures of the different macromolecules we see proteins you can have the

highest number of structures 118,000 and about 3000 nucleic acids and the complexes about 6000 this includes both protein-RNA and protein-DNA complexes.

Now, you can use this PDB to get various information you can get the collective information regarding the proteins from same families or you can create non redundant data sets or you can get the data with any specific resolution for example, high resolution structures or any specific organisms and so on.

So, if you look at it the any specific code, for example if you are interested in one protein. So, I put 2LZM in the search option right you can see the search option here right we have put the 2LZM here right and then if you click on this go right then you will get the first page with 2LZM.

(Refer Slide Time: 16:45)



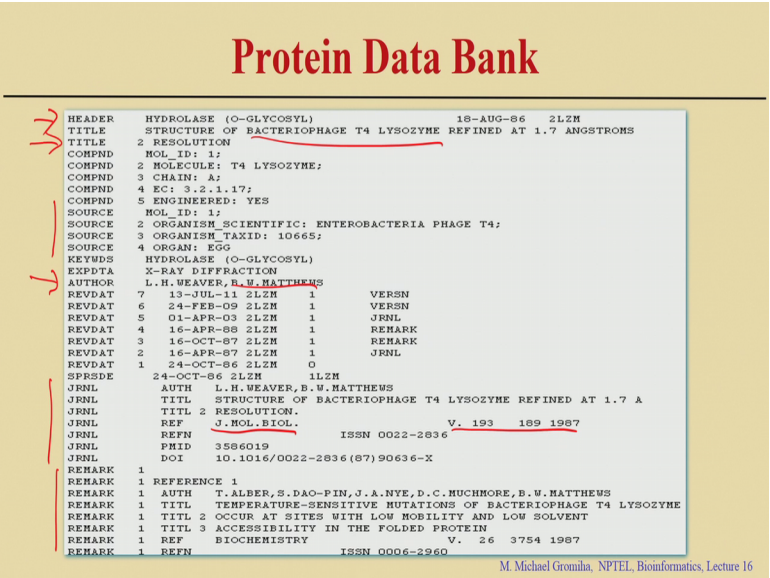
So, 2LZM tells this is bacteriophage T4 lysozyme. So, you get the abstract of that particular protein. So, here left side you can see the structure this is structure of the T4 lysozyme and you can see several options you can see here to display the file or to download the files. In the download you can see the method used here you can see these X-ray crystallography and the subunits and different other details in the summary page.

Then there are several other links for example, we can see the 3D view or the annotations sequence similarities or sequence information or the structure information experiments and so on. See when in the first page you can have different options to see get the

sequences and structures if you see here this is the FASTA sequence and I can also use the data for the full proteins, either we get the header information without coordinates or you can get the full coordinates. If you click on the FASTA sequence they will get the FASTA sequence here this is the 2LZM right this is the FASTA format right and here you start the amino acid sequence from here to here. 164 amino acid residues right you can get the sequence in FASTA format.

If you want the all the details how the particular protein just to click on the PDB format right. So, when you click on the PDB format, you can get to the information.

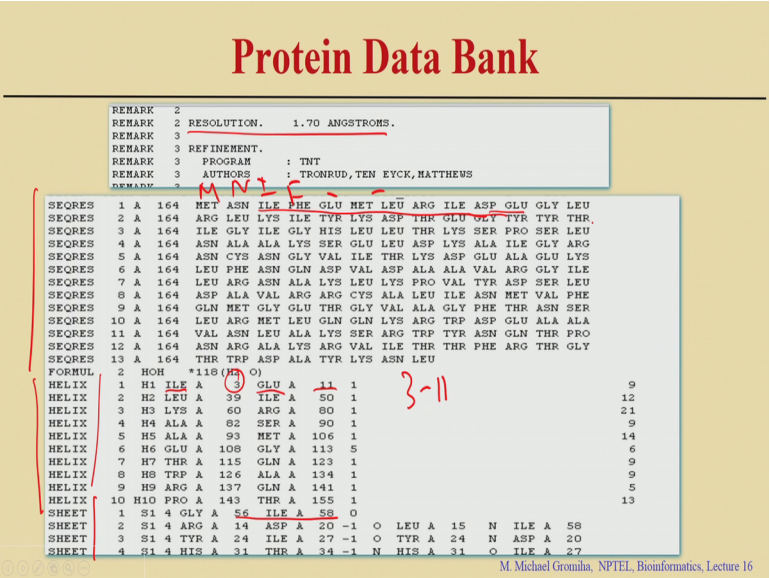
(Refer Slide Time: 18:16)



You start with a title and header this is structure of this bacteriophage T4 lysozyme and the get the more information regarding the source and you can see the authors right BW Matthews he is the one who solved the structure and he give the reference journal articles, can be published in this article right. Earlier we discussed about the Pubmed, here you can see the full reference for this structure right. Then you can see remarks you can see the several references and there are any specific remarks about the particular protein.

So, if you go down to this page right, you can see the resolution how for you can get the information about the 1.70 resolution here. So, what is the meaning of resolution?

(Refer Slide Time: 18:55)



Student: The minimum resolution accuracy.

Yeah, accuracy we can get up to this 1.7 Å accuracy level you can get the

Student: Atom

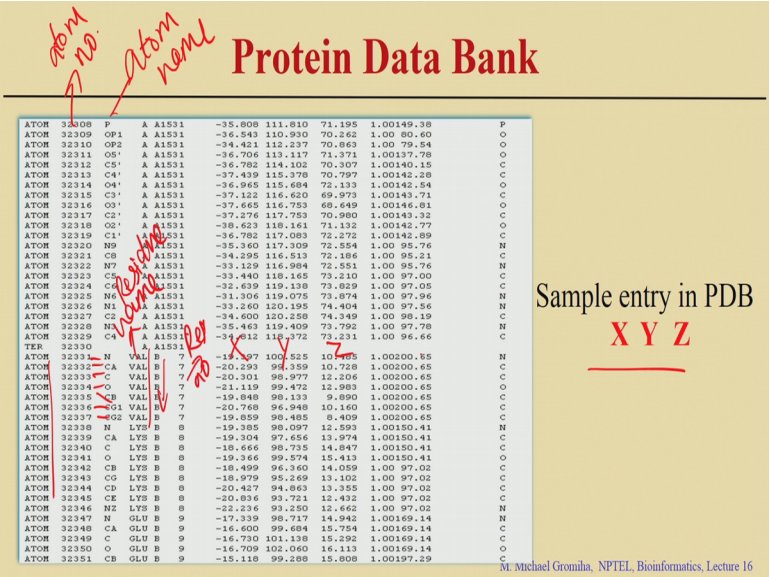
Accurate position of this each atoms right. So, earlier days we get with the high with lower resolution say 3 Å 5 Å. So, in this case the electron density maps are not so clear. So, the position of if the coordinates are also not sure right then we get more and more sophisticated instruments. Now, you can get very clear picture. So, even up to 1 Å resolution you can get the quality and the position of all the X Y Z coordinates right very reliably.

So, after that we go down you scroll down the PDB structure right. So, you can see the residue the sequence information. So, these are the sequence information. Here if the sequence are given at three little code right this is M, this is N, this I, this F and so on. And then here then we give the data about the secondary structures and if you look into the different PDB structures some of them they contain the secondary structures some of them they do not have this information. And second aspect is the end residues are not very clearly given in most of these PDB structures.

So, it contains several helices for example, isoleucine at position number 3 to glutamine acid position number 11, so 3 to 11. So, this can form a helix right from here to here right

this helical segment right. Then we can see these strands. So, there would several helices here this would the helix and you can see this the strand here right this is the segment right for example, is 56 to 68. So, this is the isoleucine glycine 56 to isoleucine 68 right this can form a beta strands here. There are 13 residues in one sequence they give. So, currently you can see the location of the secondary structure.

(Refer Slide Time: 20:51)



So, we go down we scroll down the PDB coordinates. Now, the next important thing this is the important aspect of the 3D structures. What is the important information we get from the 3D structures?

Student: 3D coordinates.

Coordinates right you need the X Y Z coordinates this is the X, Y and Z coordinates again this is atom right for I each atoms in the proteins or the nucleic acids. So, what a normal atoms in the protein? Nitrogen, hydrogen, oxygen.

Student: Sulfur.

Sulfur right. So, you can see the N C O right and then you can see this atom right N Cα CO this are the main chain and Cβ, Cγ, Cγ1, Cγ2 this side chain for this residue value.

For each atoms they give the coordinates here this is the atom number. So, this atom name, this is the atom name this is the atom number. So, here you can see the residue

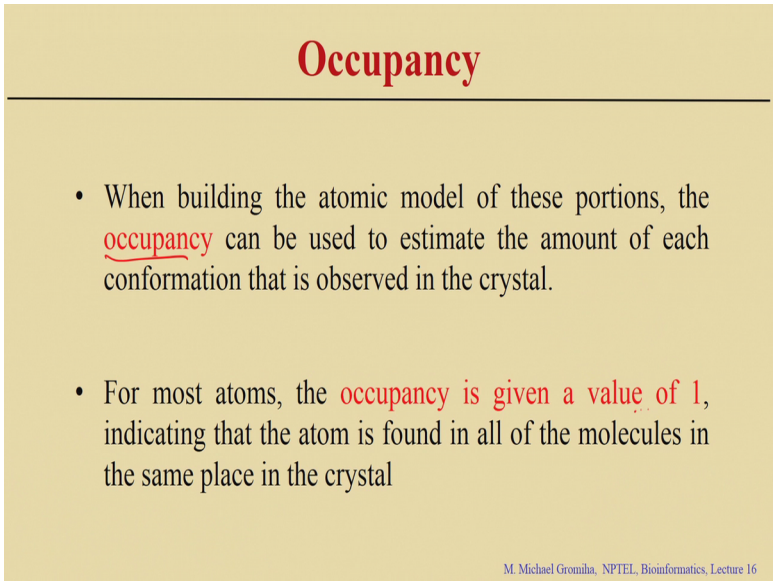
name. Now, this is the chain, chain information, this A chain or B chain this is the residue a number and then you have the coordinates the X Y Z coordinates and they give you the occupancy and the temperature factor right that I will explain very soon.

So, if we have these PDB coordinates right they give all the information regarding each atom the atom name, atom number, residue name, residue number and a chain information as well as the coordinates X Y and Z coordinates. So, they also have some water molecules and some ions also any other ligands if they are present they give as atoms heteroatoms. So, you can distinguish the atoms and heteroatoms right whether these belongs to protein coordinates or it is a water molecules are any other ions are the ligands right they used to crystallize that cocrystallize these particular protein.

Now, I discussed earlier. So, we have the XYZ coordinates right and then you can see this is the occupancy and they explain what is the occupancy which are given in the protein structure. If you look into the here this time occupancy most of the cases it is one and some cases you see there is a different numbers. What is the meaning of the different numbers and why this is one right I will now, just I will explain.

So, what is a occupancy? If you look in to these micro molecular structures micro molecule crystal there composed of many individual molecules to packed together in the symmetrical arrangement.

(Refer Slide Time: 23:18)

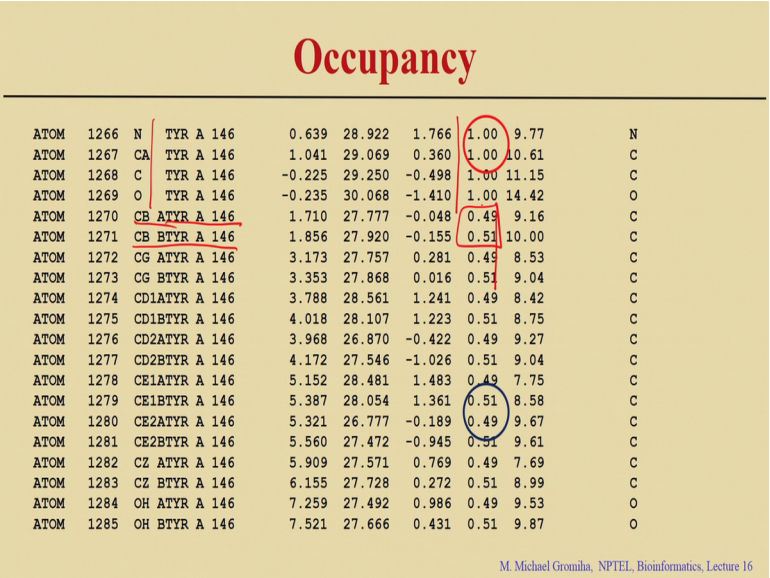


So, if you look into these structures the side chain of the surface may just wag back and forth right between several conformations. In this case a substrate can bind a different orientations either a orientation X and orientation Y right and the active site or the metal ion may be bound only few of the molecules. In this case when build the atomic model they give the term occupancy to see how much each conformation observer in the crystal whether only one conformation is observed in the crystal for each atom or this atom can have different confirmations. If the occupancy is 1 what is the meaning of that?

Student: 1

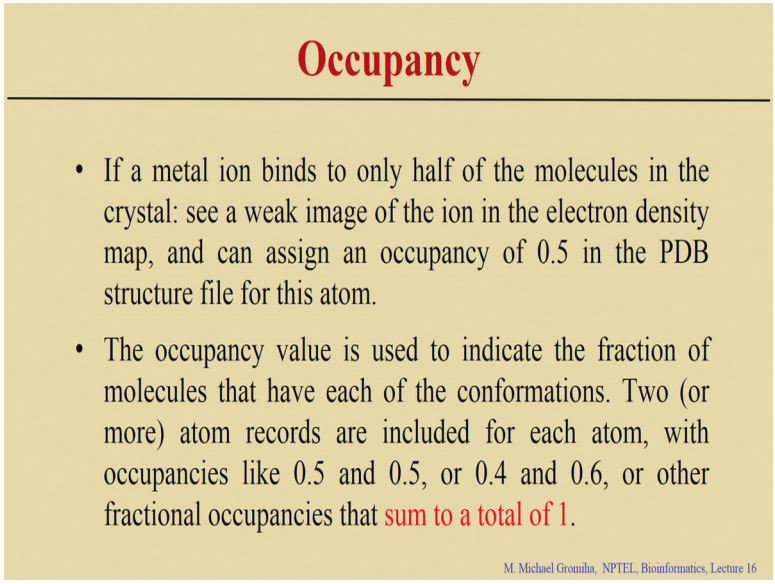
So, that atom is found in all the molecules the same place the crystal. So, there is no deviation. So, all the molecules right this atom is position in the same position if the value is 1.

(Refer Slide Time: 24:15)



In some cases this is there is a example for 1. So, if all this in this cases you can see the occupancy is 1. So, the all these atoms right these atoms in this residue right occupy the same position in all the molecules. So, this is the why the occupancy is 1. In some cases it is different say this one this is the tyrosine 146 Cβ, see if the same β tyrosine 146. Here you can see the value this is 0.49 for one case an another case it 0.51, what is the difference, how it is 0.49 and 0.51.

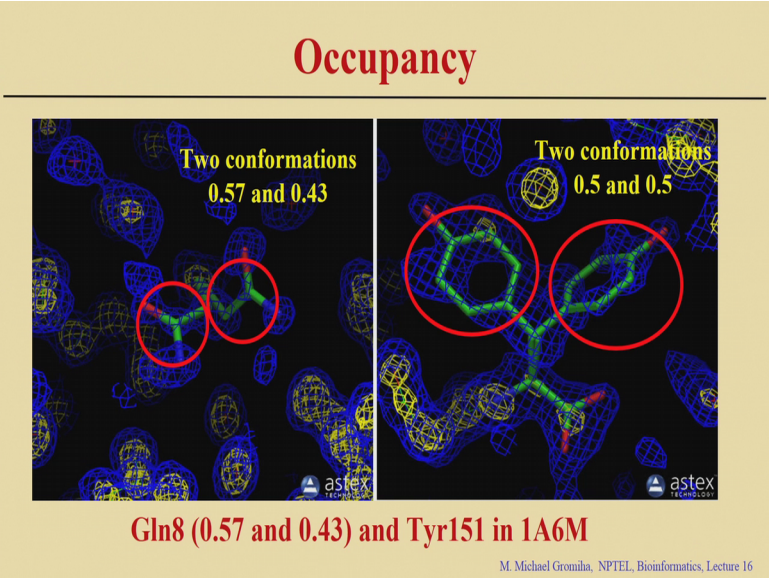
(Refer Slide Time: 24:44)



For example, if a metal ion binds only one half of the molecules you can see the electron density map right and based on that they assign the occupancy if the same orientation right they occupy then they put the 0.5 right in the both cases. What is the meaning of this occupancy? This will tell you the fraction of the molecules in different conformation for example, if two different conformations right the for I showed you here.

So, two different values 0.49 and 0.51, in this case this atom right the Cβ of tyrosine 146 they have two different orientations for example, it is 0.49, 0.51 and the total should be 1. So, there are different cases for example, 0.5 and 0.5 or 0.4 and 0.6 and any fraction occupancies, but total should be equal to 1. So, here I show an example.

(Refer Slide Time: 25:40)



So, this molecule, they have the different conformation this glutamine 8, this is glutamine 8, one is occupancy 0.57 and another one where occupancy of 0.43. This is another example here is a tyrosine 151 here both of them they have similar occupancy, so in this case this 0.5 and 0.5. If the same question is occupied we all the molecules then the occupancy is 1, depending upon the bindings, it may have different orientations if they have different conformation then the occupancy is split so that the total will be equal to 1.