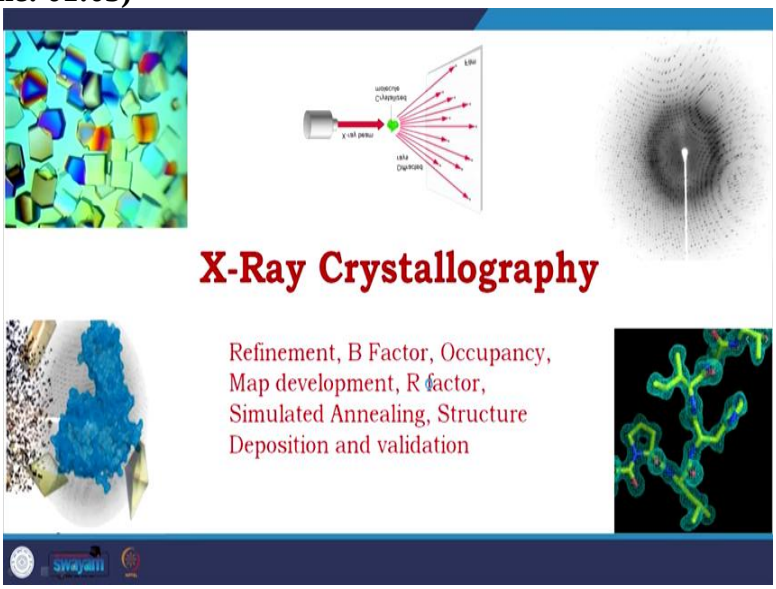
Structural Biology
Prof. Saugata Hazra
Department of Biotechnology
Indian Institute of Technology, Roorkee

Lecture - 25
X-Ray Crystallography: Refinement and Structure Deposition to PDB

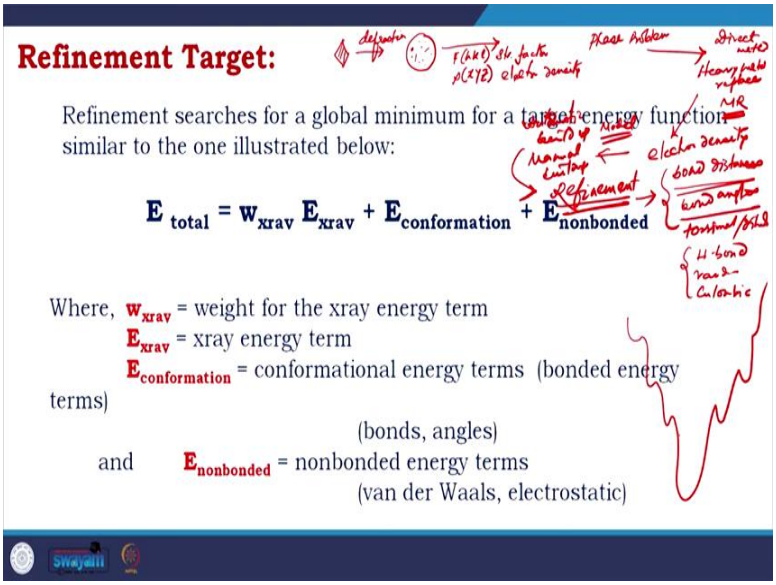
Hi, everyone. Welcome again to the structural biologic course. We are discussing structural biologic techniques, more specifically X-ray crystallography. Today, it is the last class of this long three modules of 15 classes. We have discussed many details of X-ray crystallography techniques and associated parameters, associated instruments, associated things. So, what we are going to learn today is refinement.

(Refer Slide Time: 01:03)



We will talk about refinement; we will talk about the factors we will talk about occupancy, map development, R factors, simulated annealing, structure deposition, and validation. So, from refinement, I will talk about how you get a structure and deposited to be ready to go for publication.

(Refer Slide Time: 01:24)



Refinement target. Before that, you have to understand what refinement is. So, if you remember, we get our crystal, you do the diffraction experiment, you got a diffraction pattern. You get down F (hkl), the structure factor from the diffraction pattern, and ρ (xyz) the electron density. But, we discussed here a big problem: a phase problem. You solve the phase problem in different ways, one by the direct method, then a heavy metal replacement, and third, but the molecular replacement is the best. So, with these, you get an electron density, and with the electron density, you start developing the model.

There is software that would help you make an automated build-up of the model. But the tricky part is there are some of the portions where the electron density is not good. So, here is what you have to do manually. So, think about it is not only about the theory. It is also about understanding the concepts of how things work. So, what we are doing is that you have electron density. You manually put the amino acids initially the backbone of the amino acid, the main chain of the amino acids, and then after that sidechain. So, by your manual estimation, you are putting the amino acid on the electron density. So, let us say automatic build-up and manual build-up of the model. So, when you are doing manual build-up, you are making natural mistakes. That is why it is manual. How to correct it? To correct it, what generally we do is called refinement.

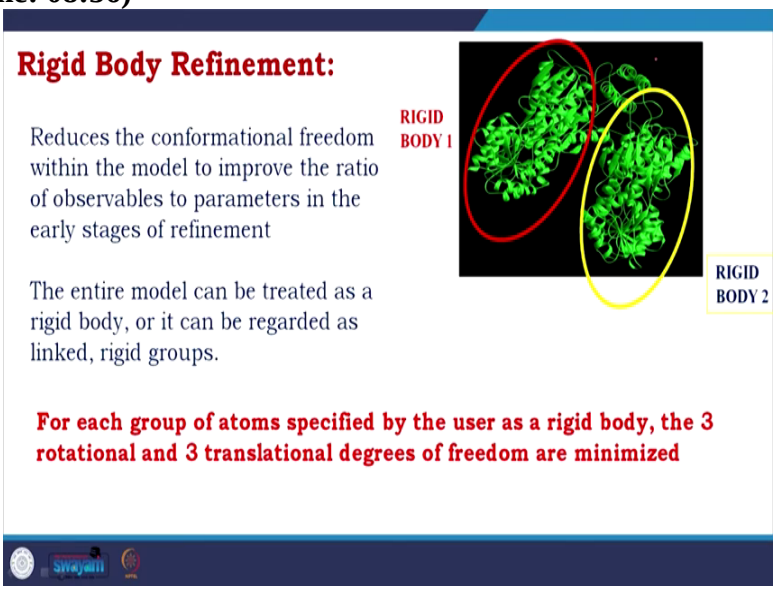
So, if you see refinements search for a global minimum for a target energy function, similar to the one illustrated

$E_{total} = W_{X\text{-}ray} E_{X\text{-}ray} + E_{conformation} + E_{non\text{-}bonded}$

Where $W_{X\text{-}ray}$ is the weight for the X-ray energy term. $E_{X\text{-}ray}$ is X-ray energy terms, E conformation is conformational energy terms, or bonded energy terms which talks about

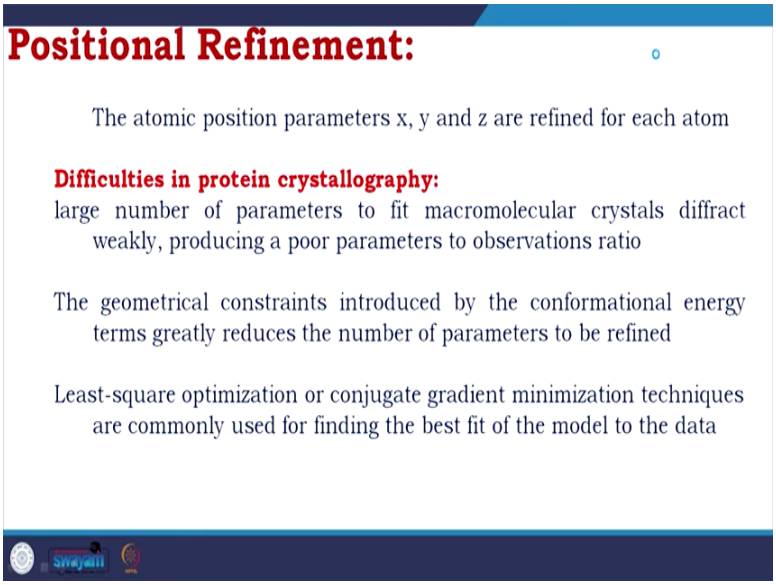
bonds, angles, and $E_{non\text{-}bonded}$ is the non-bonded energy terms talking about van der Waals and electrostatic.

(Refer Slide Time: 08:56)



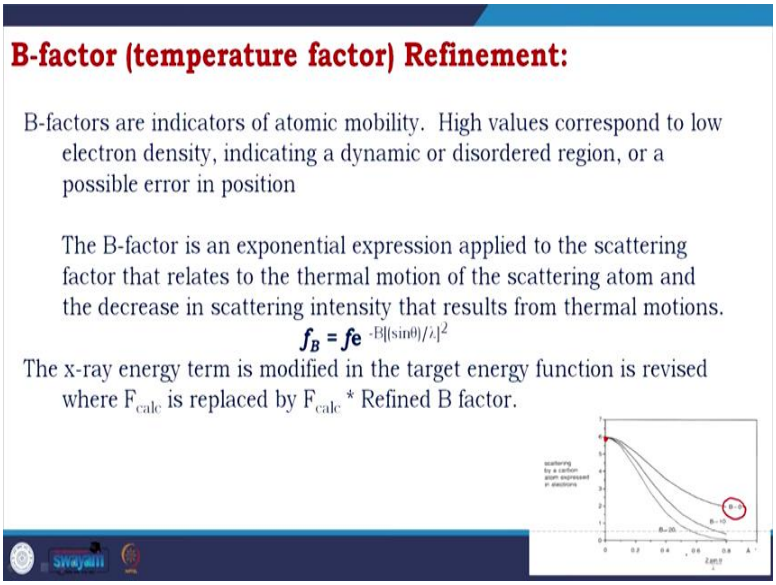
So, there are different types of refinement. I am discussing the basic ones for your understanding because this is an introductory course, as I told you multiple times. So, rigid refinement or rigid body refinement, what it does, it reduces the conformational freedom within the model to improve the ratio of observables to a parameter in the early stage of refinement. What it is talking about, suppose, you are building up fingers of these two hands. If they rotate between each other, you will get more probabilities more degrees of freedom. So, if you consider them the rigid body, you model the movement between them. That is what we are doing here. The entire model can be treated as a rigid body or regarded as linked, rigid groups. As you see, there are two different domains. One domain is considered rigid body 1, and the other is considered rigid body 2. The rotational and translational degrees of freedom are minimized for each group of atoms specified by the user as a rigid body. So, what is the difference between rigid body refinement and normal refinement in normal refinement? You allow every rotation and translation here, and you are not allowing the rotation and translation of the molecules inside here because you want to see the movement of one rigid body concerning the other rigid body.

(Refer Slide Time: 10:59)



Coming to the positional refinement, the atomic position parameters x y and z are refined for each atom. There are difficulties in protein crystallography. Many parameters have to be fit in the macro molecular crystals, especially those that defect quickly, producing a poor parameter to observation ratio. The geometrical constraint introduced by the conformational energy terms, which are bonded terms, greatly reduces the number of parameters to be refined. Least-square optimization or conjugate gradient minimization techniques are commonly used to find the model's best fit to the data. So, when we are doing a rigid body, it is relatively easier because we are not considering all the motions. Here we take least-square optimization and conjugate gradient because they are generally less computer calculation intensive.

(Refer Slide Time: 12:17)



B-factor or temperature factor refinement indicates that B-factors are atomic mobility indicators. High values correspond to low electron density, indicating a dynamic or disordered region or a possible error in position. In general, a high B-factor says flexibility or disorders in the structure if you consider B-factor. Still, it might also be corresponding to low electron density for a small molecule. It will never happen because the molecule is exposed X-rays diffracting you get the diffraction data to calculate you get a high B-factor, you could be sure that this is because the molecules are moving. Low electron density could come from other factors that could affect the B-factor. The B-factor is an exponential expression applied to the scattering factor that relates to the thermal motion of the scattering atom and the decrease in scattering intensity that results from the thermal motion. So, it is correlating to the thermal motion. That is why it is called proportional or correlated to flexibility.

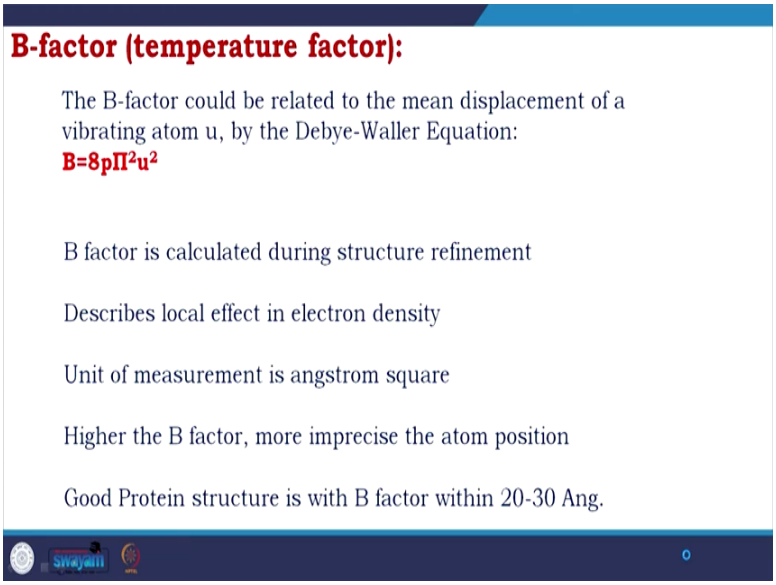
The formula is

$f_B = f_e^{[- B \sin\theta / \lambda]2}$

F_B is the modified scattering factor, with B the temperature effect. The X-ray energy term is modified in the target energy function is revised where F_{calc} is the calculated structure factor is replaced by F_{calc*} star with the refined B-factor.

So, if you look at here, scattering by a carbon atom explicitly electrons because carbon has 6 electrons, you see that it starts from here. Now, when you put these with different B - factor B = 0 B = 10 B = 20 the change in the angstrom 2 $\sin\theta / \lambda$ value.

(Refer Slide Time: 15:34)

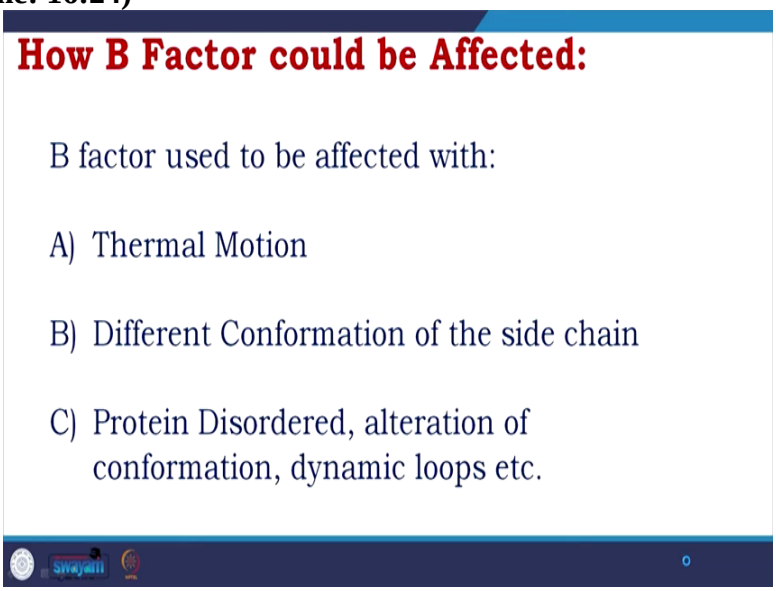


The B-factor could be related to the mean displacement of a vibrating atom u, by the Debye-Waller equation, which is

$B = 8 p \pi^2 u^2$

B-factor is calculated during structure refinement, describes local effects in electron density. If electron density is low, you will find a high B-factor. Unit of measurement is angstrom square, higher the B-factor more imprecise the atomic position, good protein structure is B-factor within 20 to 30 Ang.

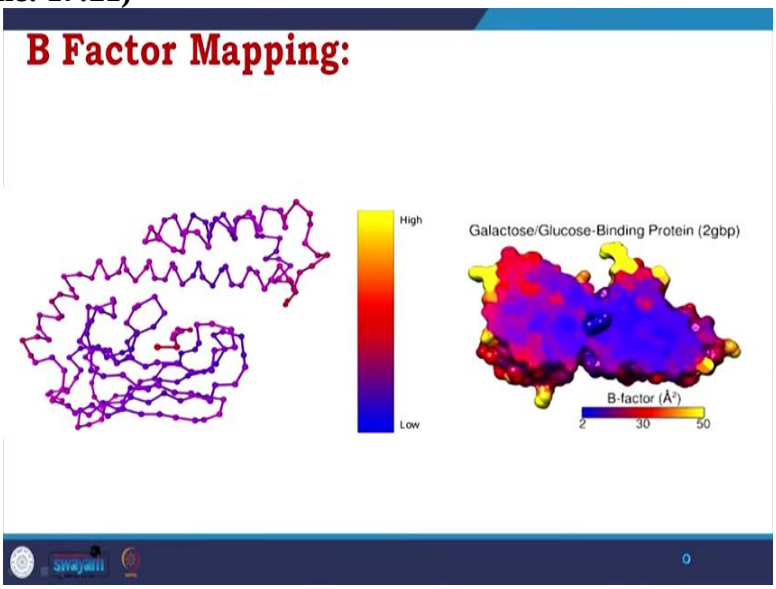
(Refer Slide Time: 16:24)



How B-factor could be affected: B-factor used to be affected with thermal motion, which I have already discussed, because, with increasing thermal motion there is flexibility when there is flexibility, the X-ray is hitting one time this position other time this position so, resulting in low electron density or poor electron density. Different conformation of the sidechain if the sidechain conformation were different, it would automatically fluctuate

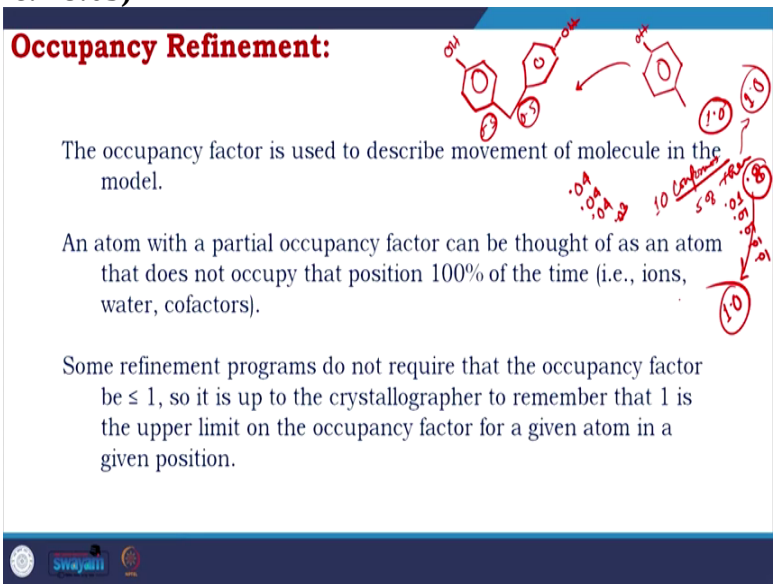
between a conformer and b. Hence, having a motion protein disordered pair put in is not getting you to know a very ordered structure like helix, beta sheet. It has a general loop that is also flexible, alteration of conformation where there are changing and dynamic loops.

(Refer Slide Time: 17:21)



You will see the B factor map, and the values from blue to yellow is orange showing low to high B factor. So, if you see in the core, the B factor is low, whereas, in the surface, you generally see higher B factor because the amino acid side chains are not interacting with the core in the surface. So, they are free. So, they are freely moving because they are creating high motion. If there is the disorder portion, they are also moving. So, it shows higher B factor.

(Refer Slide Time: 18:05)



The occupancy factor is used to describe the movement of molecules in the model. An atom with a partial occupancy factor can be thought of as an atom that does not occupy that position 100% of the time. Occupancy is a very interesting subject. Considering occupancy,

you could see the alternative conformation that means you could map the movement of the atoms.

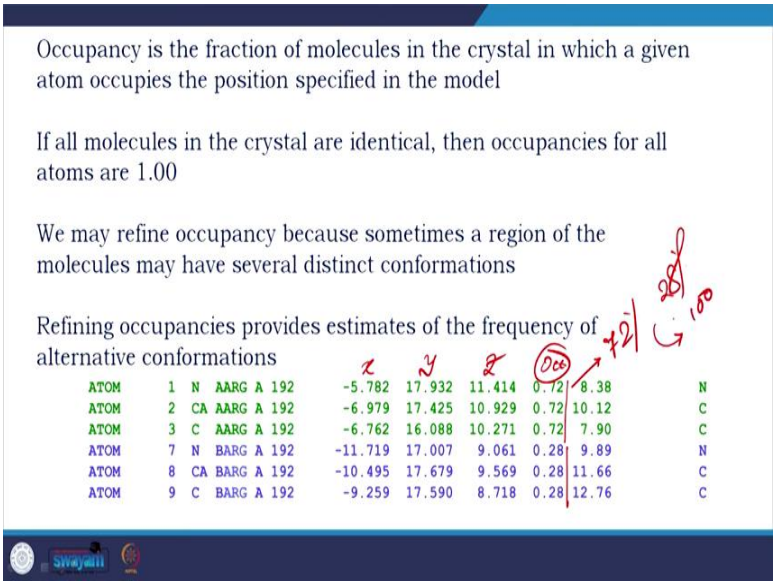
Though the overall structure and results in the PDB are static, suppose we are talking about a tyrosine. If the tyrosine has one conformer, the occupancy would be 1.0. But if the tyrosine has two conformers, the occupancy would probably be 0.5, 0.5.

So, it should consider the conformer up to the level of intensity it is providing what I mean by that, let us say there is tyrosine. It had ten conformers, 5 of them would contribute points let us say 01% each and then there are 0.95, one is having let us say 0.9 or 0.8 and other four are having 04, 04, 04 and 03. So, 10 conformers 4, 8, 12, 15, 20, 0.8 100, 1. I mean, you would benefit from knowing the contribution of conformations.

Because some of the confirmation that does not contribute higher stability might play an important role in the development of that transition state intermediates and all things like that, still, the general trend of the software is to show that this confirmation as 100% contributes.

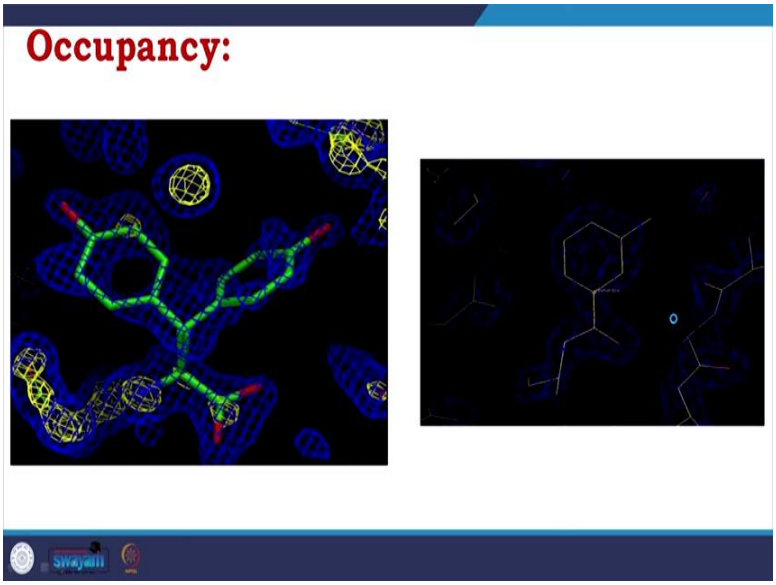
Some refinement programs do not require the occupancy factor to be less than equal to 1. So, it is up to the crystallographer to remember that one is the upper limit on the occupancy factor for a given atom in a given position. We have already discussed that.

(Refer Slide Time: 22:34)



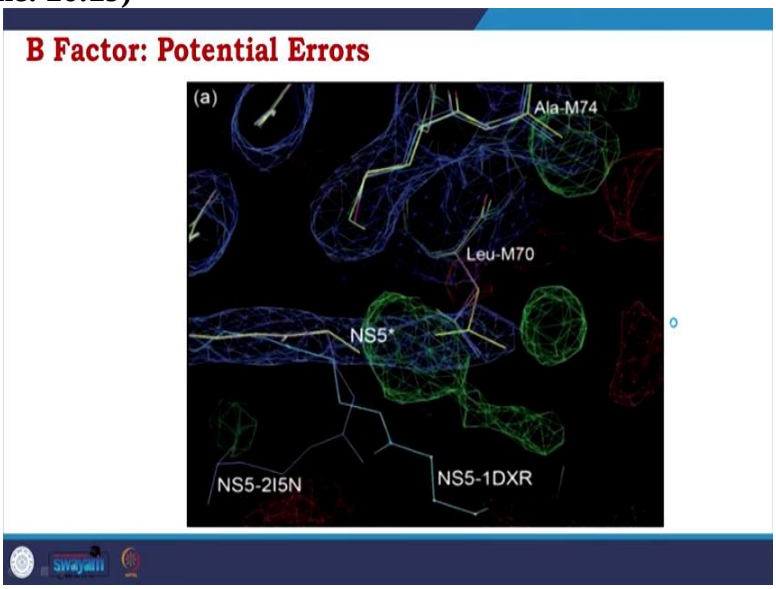
Occupancy is the fraction of molecules in the crystal in which a given atom occupies the position specified in the model. If all molecules in the crystal are identical, then occupancies for all atoms are 1.00. We may refine occupancy because sometimes, a region of the molecule may have several distinct conformations. Refining occupancies estimate the frequency of alternative conformations if you see them. So, before showing this data, I want to tell you one thing: we are discussing structural biology and structural biology techniques, but I have to introduce you to PDB. What is PDB? PDB is a format it is called protein databank where all the biological macromolecules whose 3D structures are solved are deposited. Because whatever we are doing, the output would come in the dot pdb format, which is a format where you get the x, y, z coordinates. So, I am not going into details because I am spending a few classes on this, but you will see that the x, y, and z are there, and these are the occupancy. So you see, this is a side chain of the arginine, and there is for one confirmation 72%, which means 0.72 other 0.28 28% confirmation is 100.

(Refer Slide Time: 26:14)



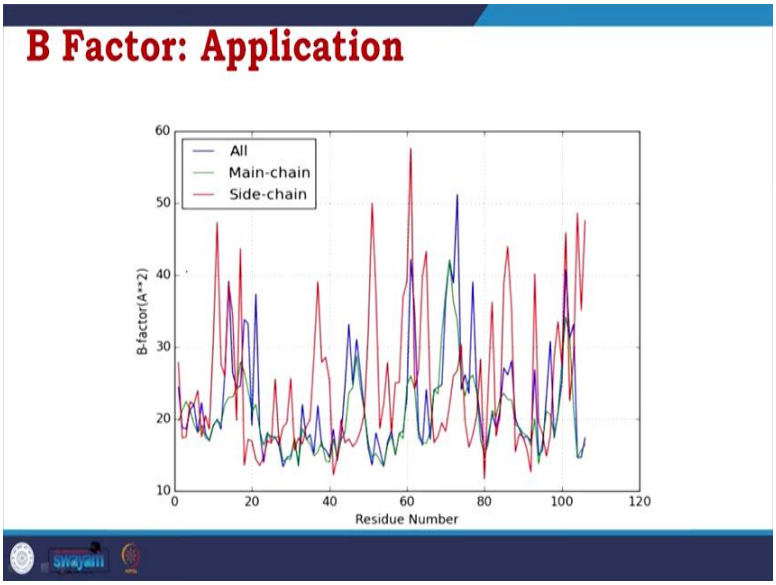
See, two confirmations showing occupancy here you will see, so, one is having 0.51 is 7.5 or altering here also you see the difference in the positions they occupy.

(Refer Slide Time: 26:29)



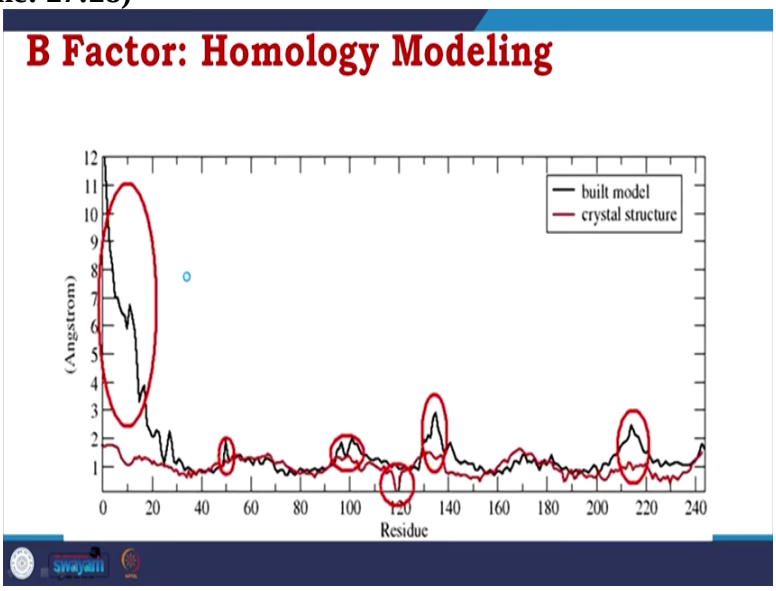
And also, you could identify potential errors here, and you click any check the occupancies, you could correct the maps.

(Refer Slide Time: 26:48)



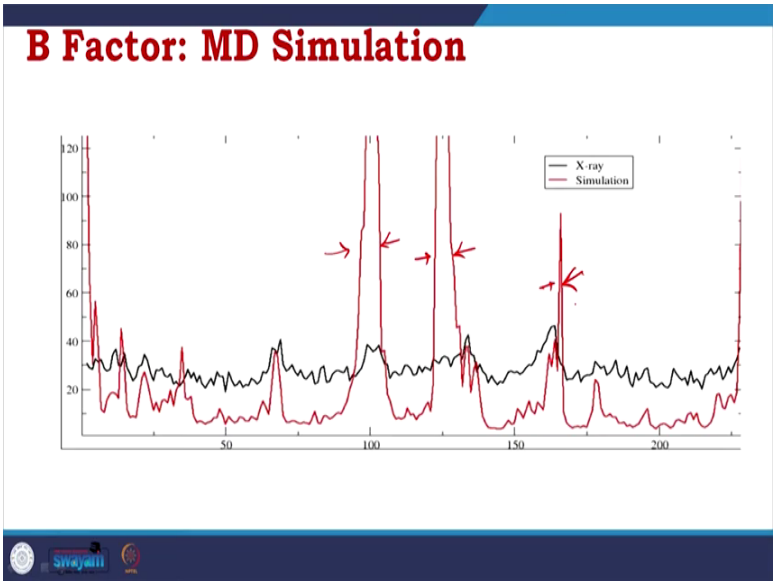
Application of B factor: If you plot the whole protein main-chain, sidechain, you will see for sure more movement on the side chain. That is why you see that B factor residue number you see higher plot higher B factor for the side chains because main-chain is forming the core of the structure, they are more stable definitely than the sidechains.

(Refer Slide Time: 27:28)



This is a comparison of a crystal structure with a built model. Even in most cases, you see, they do agree with the crystal structure, but at the initial few residues. They saw this orderliness from the homology model structure by comparing the B factor. It is one way to check the validity of a homology model if you have corresponding crystal structures present.

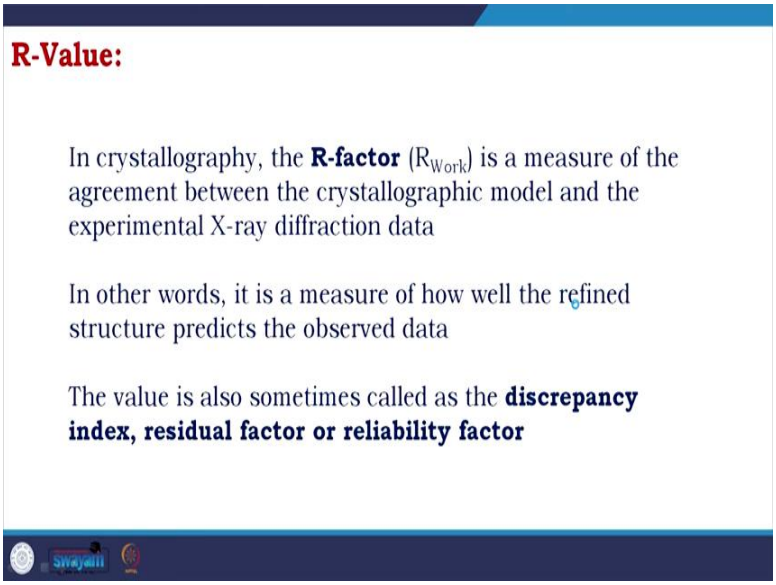
(Refer Slide Time: 28:07)



More interestingly, in simulation, as I talked about, in the simulation, we have a force field and develop a force field parameter depending on what type of factors are influencing the biological macromolecule. For example, for protein, we take bonded parameters: bond distance, bond angle, bond distance, bond stretching, and angle bond bending and torsion. For non-bonded, it is columbic interaction and Van der Wall interaction and hydrogen bonds.

When you do that, you find that which residues like here, nearly 100, you get a very flexible portion. There are probably loops, and you could get to know the dynamics of the loop by applying MD simulation. But comparing the B factor from the MD simulation and crystal structure, you could easily get that difference.

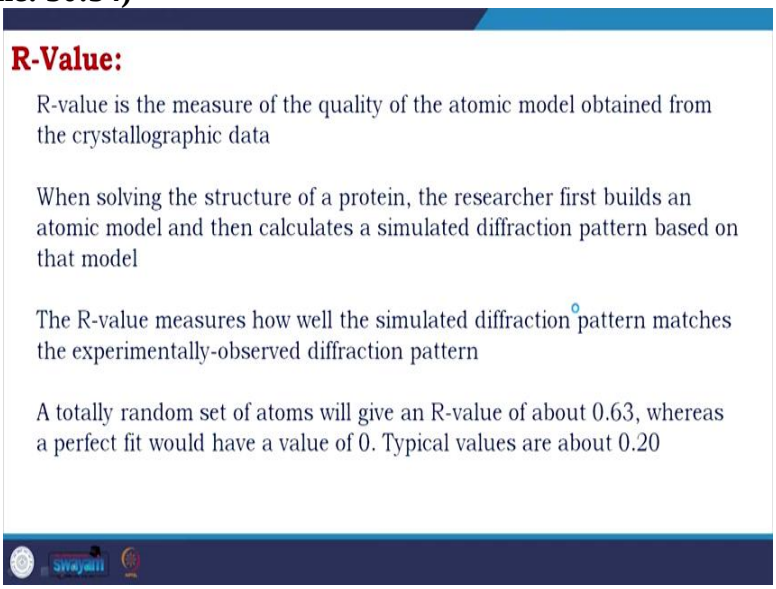
(Refer Slide Time: 29:16)



Now I will talk about a very interesting validation called R-value in crystallography. The R-factor or R work measures the agreement between the crystallographic model and the experimental X-ray diffraction data. You get the crystal, you diffract, and you get the diffraction data diffraction patterns, and you develop a model. Is there any disagreement between is there any error happened?

R is a good measure of that. In other words, it is a measure of how the refined structures will predict the observed data you get data, and you have a defined structure, how well they are coordinated could be estimated from the R-value. The R-value, which they call R-factor, is also called discrepancy index residual factor or reliability factor.

(Refer Slide Time: 30:34)

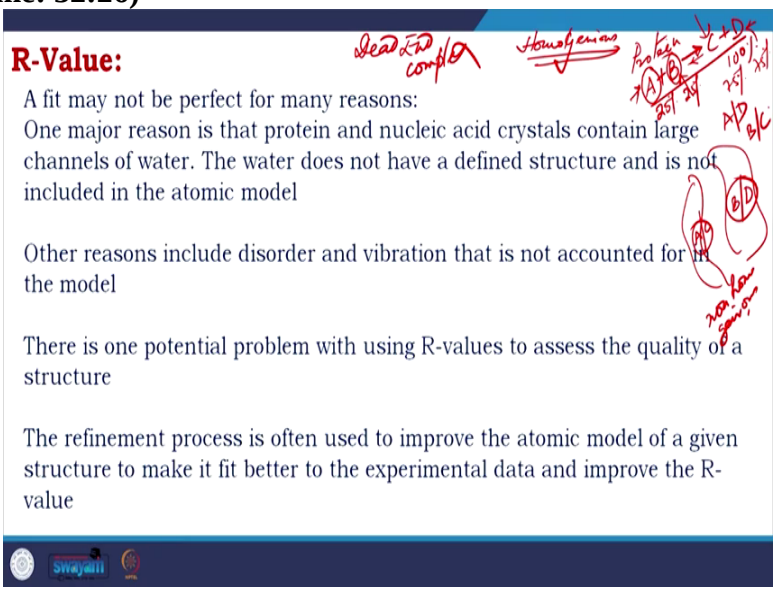


R-value measures the quality of the atomic model obtained from the crystallographic data. When solving the protein structure, the researcher first builds an atomic model and then calculates a simulated diffraction pattern based on the model. The R-value measures how

well the simulated diffraction pattern matches the experimentally observed diffraction pattern.

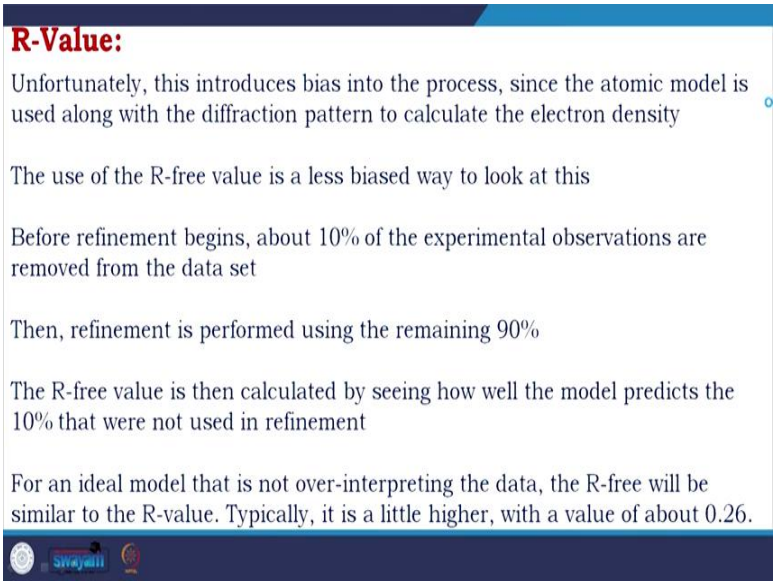
A totally random set of atoms will give an R-value of about 0.63. So, a totally random set of atoms will give an R-value of about 0.63, whereas if perfect fit would have been a value of 0.

(Refer Slide Time: 32:26)



In R-value, a fit may not be perfect for many reasons. One major reason is that protein-nucleic acid crystals contain large channels of water. The water does not have a defined structure and is not included in the atomic model. Other reasons include disorder and vibration not accounted for in the model. One potential problem with using R-values to assess the quality of a structure is the biasness. The refinement process is often used to improve the atomic model of a given structure to fit the experimental data and improve the R-value.

(Refer Slide Time: 42:11)



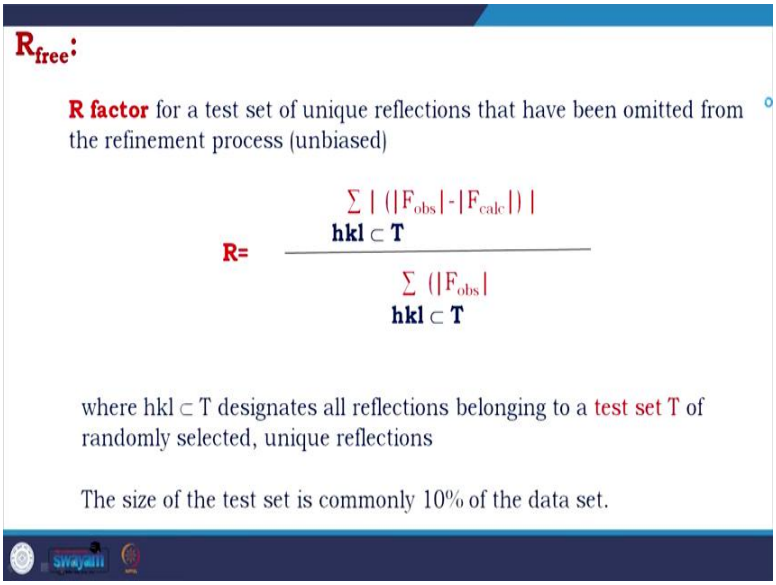
Unfortunately, this introduces bias into the process since the atomic model is used along with the diffraction pattern to calculate the electron density. R-free value is the less biased way to look at this. Before the refinement begins, about 10% of the experimental observations are removed from the data set. Suppose in your school, and there are 1000 students. You want to develop a machine learning program to find a few features. Now, if you collect data from 1000 students and develop a model, all of the student's data is inside the model. So, if you use that model to predict anything about these 1000 students, it would always be perfect. That is called the biasness because you have already taken everyone's data. So what to do? To develop a better program, you need virgin data. What do I mean by virgin data?

By virgin data, I mean your school have 1000 student. If you include 900 students' data and keep 100 students' data separate to make a model using the data from the 900 students and apply this AI-based model the machine learning based model to these 100 students and you get good results, you know your program is good. Similarly, what do we do? We get the entire R-value, and we have taken 10% of R before the refinements.

So, we get the experimental observations of the frames, and we take out 10% of the data, then refinement is performed using the remaining 90% data. The R-free value is then calculated by seeing how well the model predicts the 10% that were not used in the refinement for an ideal model, not over-interpreting the data that R-free will be similar to our value typically, it is a little higher. So, if you say 0.2 is the range of R-work, 0.26 is the range of R-free.

(Refer Slide Time: 45:53)

(Refer Slide Time: 47:18)

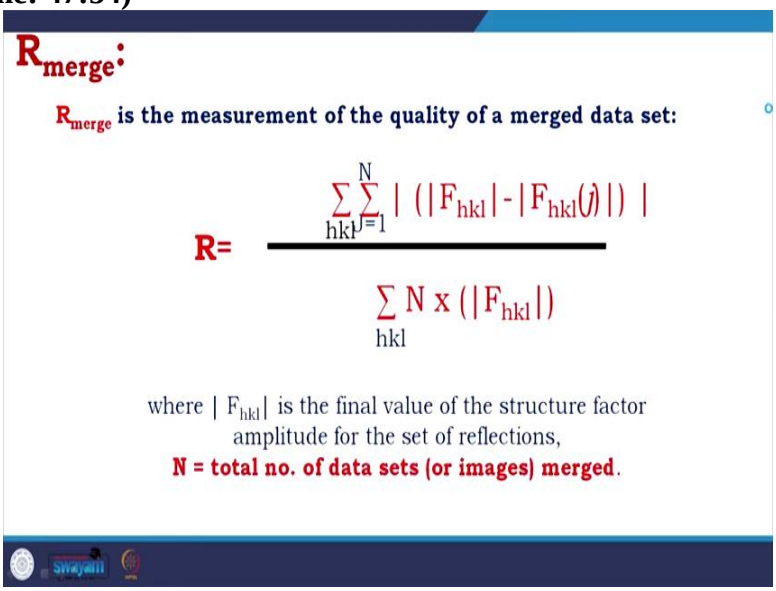


R-factor for a test set of unique reflections that have been omitted from the refinement process

R = hkl to T with $\sum \left| \left(\left| F_{obs} \right| - \left| F_{calc} \right| \right) \right| / \sum \left(\left| F_{obs} \right| \right)$

where hkl to T designates all reflections belonging to a test set T of randomly selected unique reflections, the site of the test set is commonly 10% of the data set.

(Refer Slide Time: 47:54)

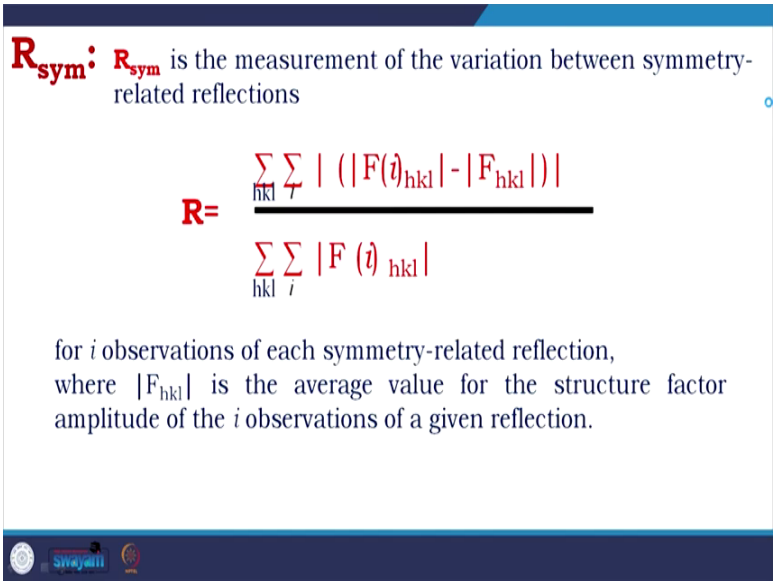


R merge is the measurement of the quality of a merge data set

R = $\sum \left| \left(\left| F_{hkl} \right| - \left| F_{hkl}(j) \right| \right) \right| / \sum N \times \left(\left| F_{hkl} \right| \right)$

Summation j = 1 to N, absolute F hkl is the final value of the structure factor amplitude for the set of reflection. N is the total number of data sets as we talked about different data sets merging here.

(Refer Slide Time: 48:22)

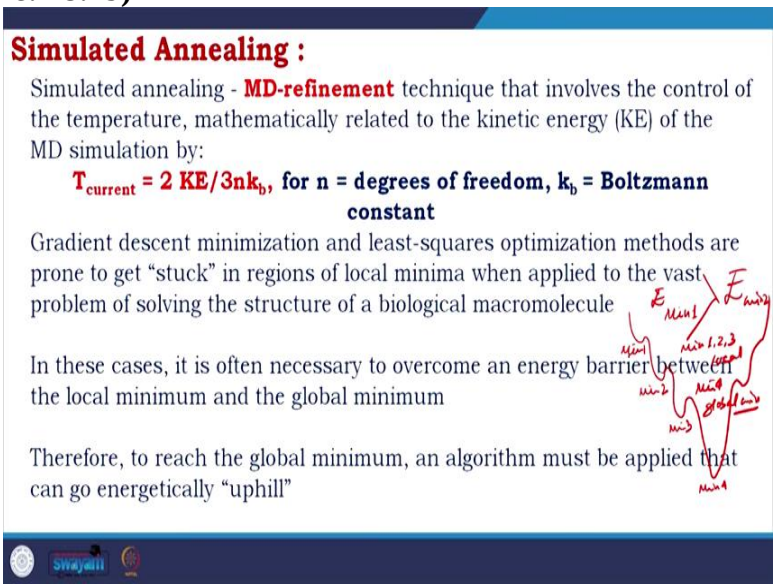


R sym as the measurement of the variation between symmetrically related reflections

$R = \sum\sum |(|F(i)_{hkl}| - |F_{hkl}|)| / \sum\sum (|F(i)_{hkl}|)$

you see summation hkl with summation i, i is an inversion center. For i observations of each symmetry-related reflection where the absolute value of F hkl is the average value for the structure factor amplitude for the i observation of a given reflection.

(Refer Slide Time: 48:49)



Simulated annealing is an MD-based refinement technique that involves the control of the temperature mathematically related to the kinetic energy of the MD simulation by

$T_{current} = 2\ KE / 3\ nk_b$

for n = degrees of freedom, k_b equal to Boltzmann constant.

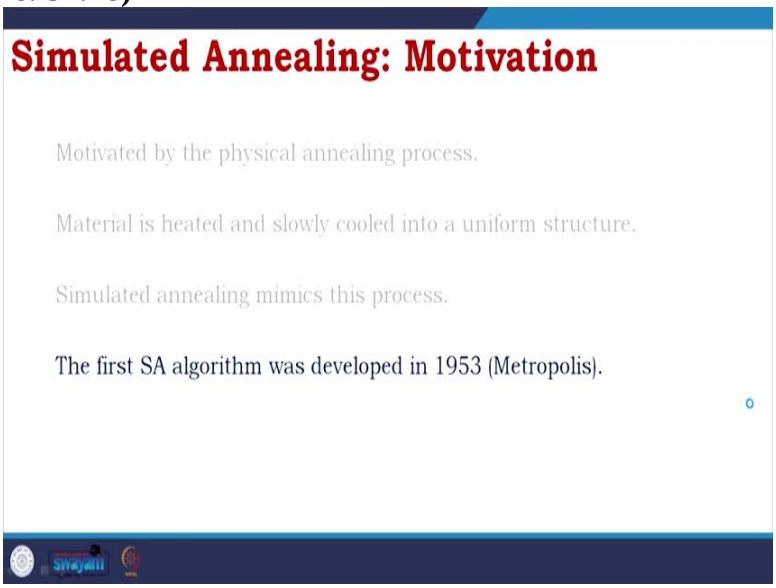
Gradient descent minimization and least-square optimization methods are prone to get stuck in regions of local minima when applied to the first problem of solving the structure of biological macromolecules. Do you understand? Let me talk about this. So, suppose we are

talking about a protein. I talked about how there are different minimas, minima 2, minima 3, minima 4, and other minima. You can easily understand that minima 1's energy is greater than minima 2.

So, minima 1, 2, 3 are local minima, and minima 4 is the global minima. Simulated annealing is the process where you have to overcome those local minimums.

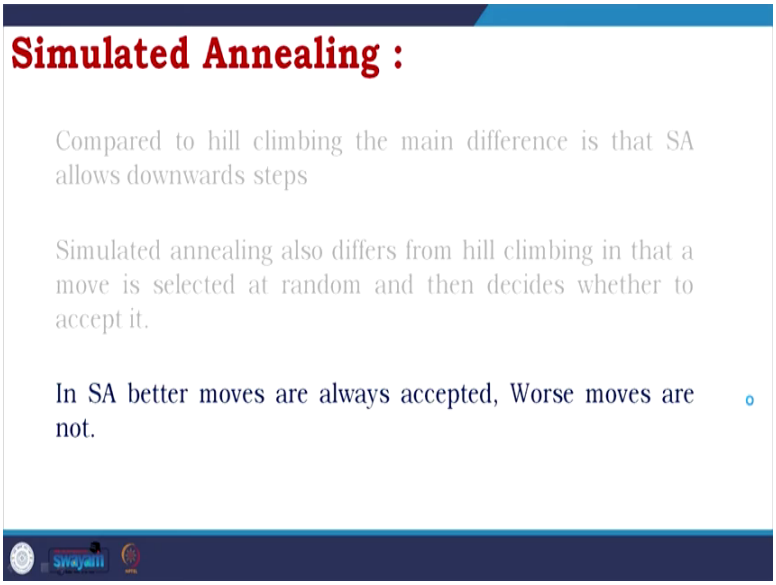
Let us check in these cases it is often necessary to overcome an energy barrier between the local minimum and the global minimum. Therefore, an algorithm must be applied that can go energetically uphill to reach the global minimum.

(Refer Slide Time: 52:28)



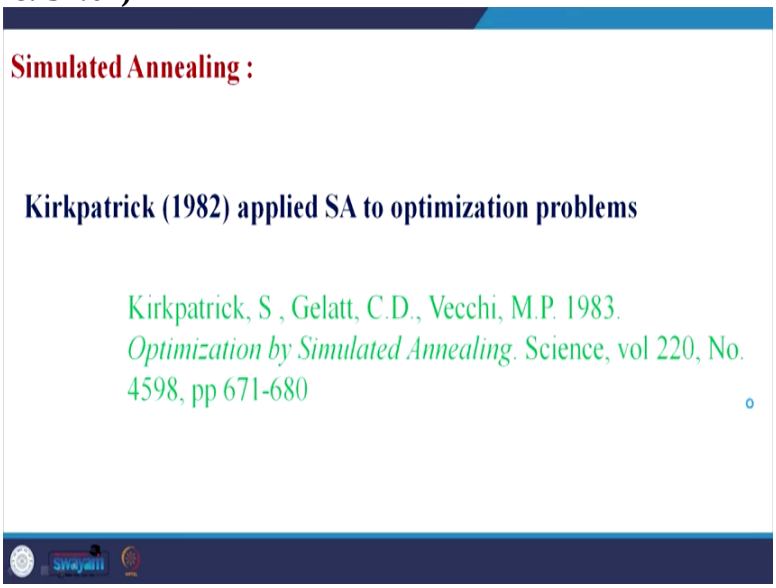
The motivation of simulated annealing by the physical annealing process means the heating process. So, it is a computational heat provider. Material is heated and slowly cooled into a uniform structure. Simulated annealing mimics this process. So, it provides heat and then goes for cooling down. The first simulation algorithm was developed in 1953 by Metropolis.

(Refer Slide Time: 53:16)



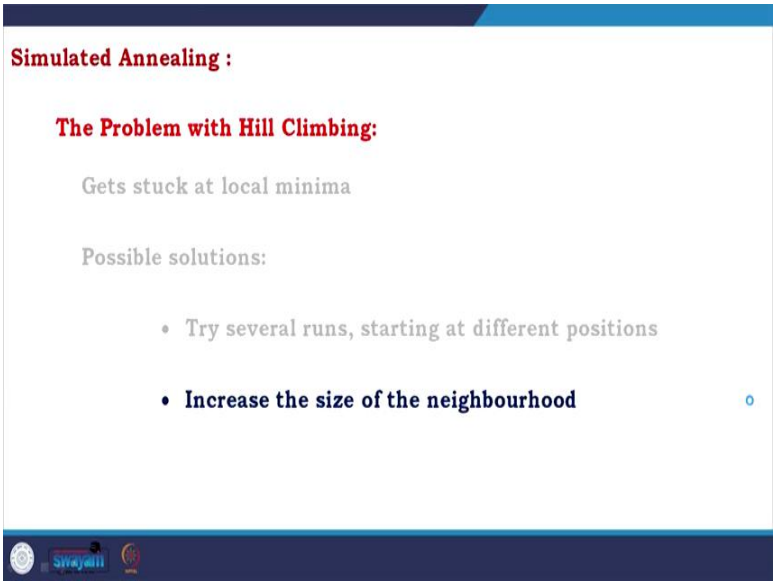
It could be compared to hill climbing. The main difference is that simulated annealing allows downward steps. Simulated annealing also differs from hill climbing in that a move is selected at random and then decides whether to accept it or not. So, there will be random moves, and then you have to go for testing. In simulated annealing, better moves are always accepted; worse moves are not.

(Refer Slide Time: 54:01)



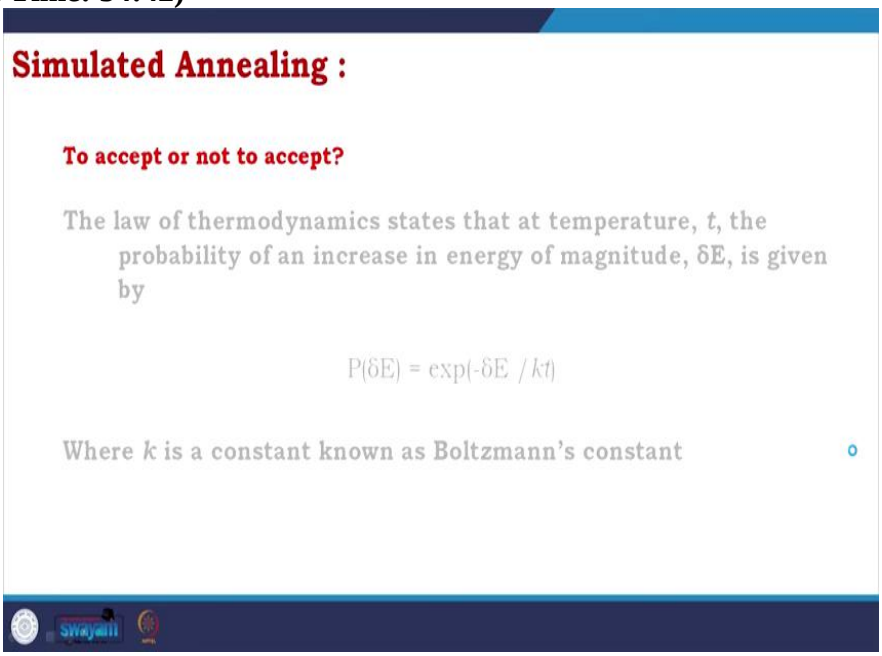
Kirkpatrick, 1982 applied simulated annealing to optimization problems. The paper is given here. It is published in Science. So, this is for your reference, you could check it.

(Refer Slide Time: 54:20)



So, the problem of hill climbing is stuck at local minima, which we talked about possible solutions. Try several runs starting at different positions; increase the size of the neighborhood.

(Refer Slide Time: 54:42)

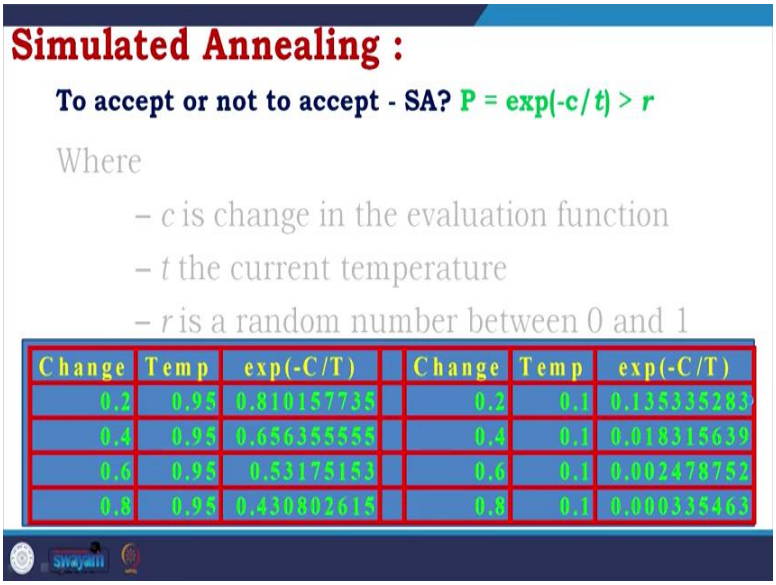


But then you have to choose the factor to accept or not to accept. The law of thermodynamics states that at temperature t, the probability of an increase in energy of magnitude ∂E is given by

$P(\partial E) = \exp(-\partial E / kt)$

where k is a constant known as Boltzmann's constant.

(Refer Slide Time: 55:16)

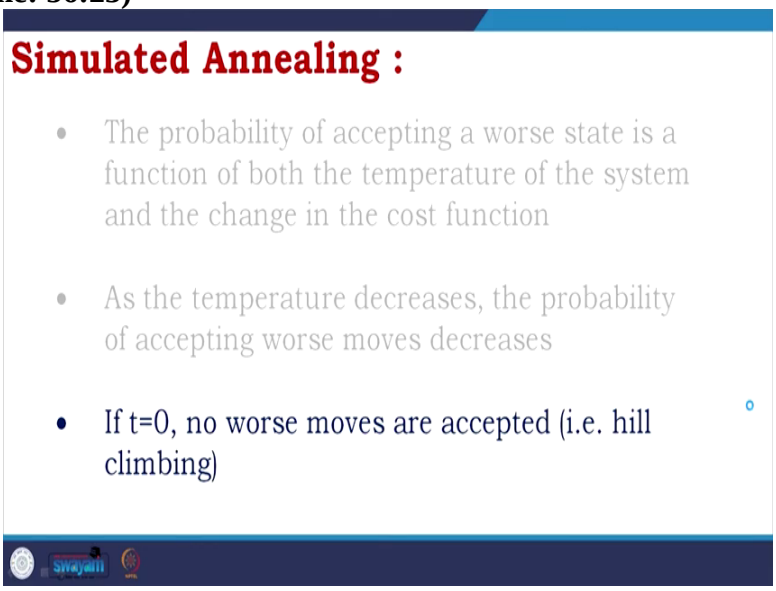


So, as I told you to accept or not to accept, simulated annealing the steps

P (probability) = exp (-c / t) > r

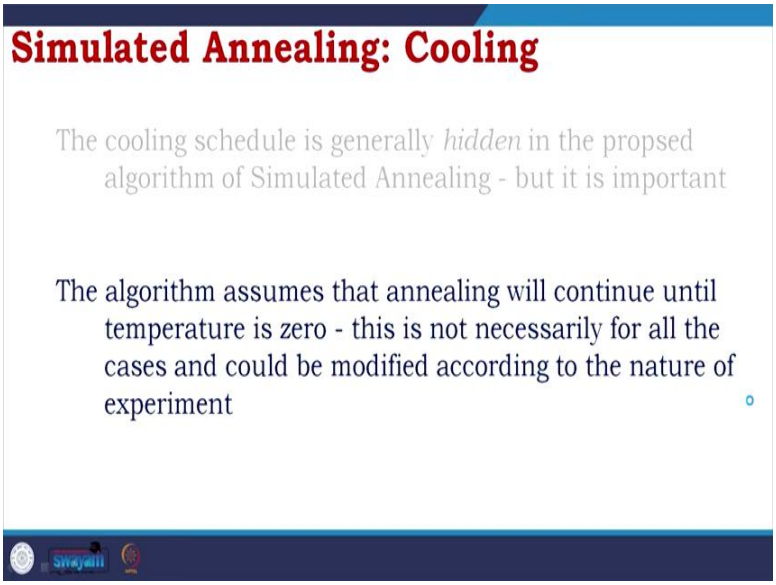
that should be the condition where c is a change in the evaluation function, t is the current temperature, r is a random number between 0 and 1.

(Refer Slide Time: 56:23)



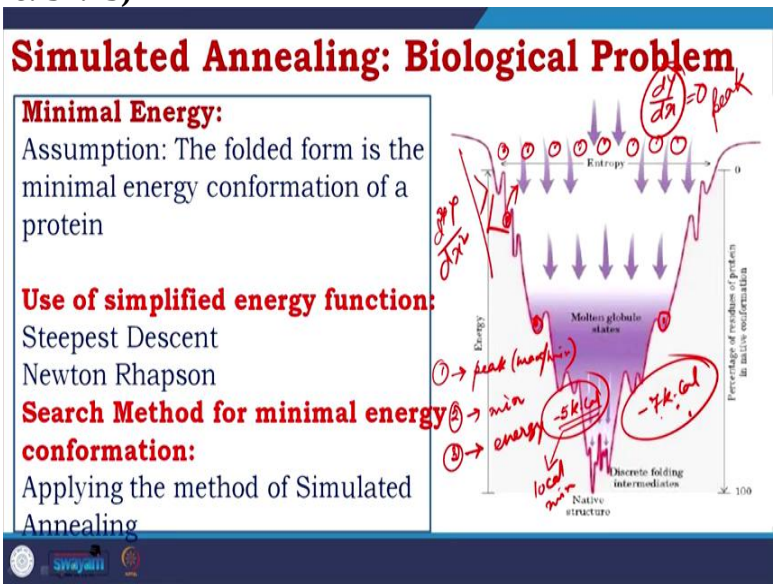
The probability of accepting a worse state is a function of both the temperature of the system and the change in the cost function. As the temperature decreases, the probability of accepting worse moves decreases. If t = 0, no worse moves are accepted. That is you have to go hill climbing.

(Refer Slide Time: 56:53)



The cooling schedule is generally hidden in the proposed algorithm of simulated annealing. Still, the algorithm must assume that annealing will continue until the temperature is 0- this is not necessarily for all the cases and could be modified according to the nature of the experimental nature of the macromolecule nature of the experimental condition you are looking for.

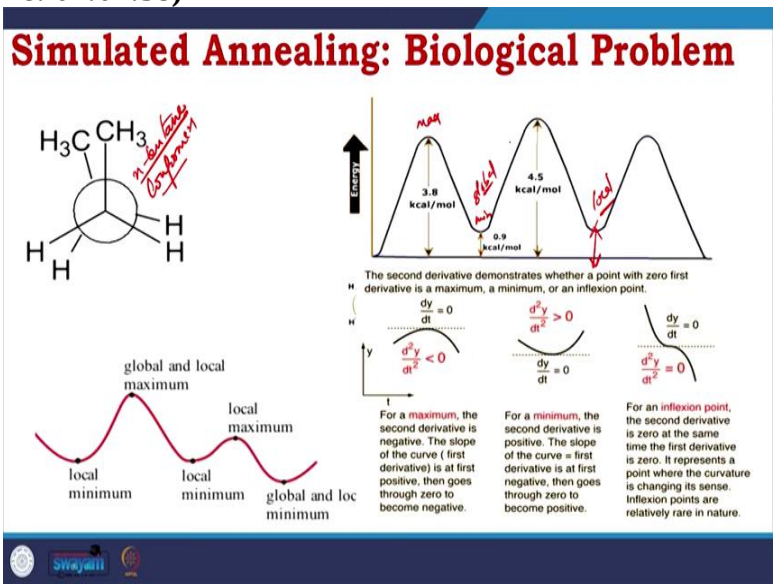
(Refer Slide Time: 57:23)



In simulated annealing, assuming the folded form is the minimum energy conformation of a protein, we use a simplified energy function and search method for minimal energy conformation. So, simplified energy functions, which are easy computer less computer-intensive steepest descent and Newton Rhapson and search method for minimal energy conformation, apply the method of simulated annealing.

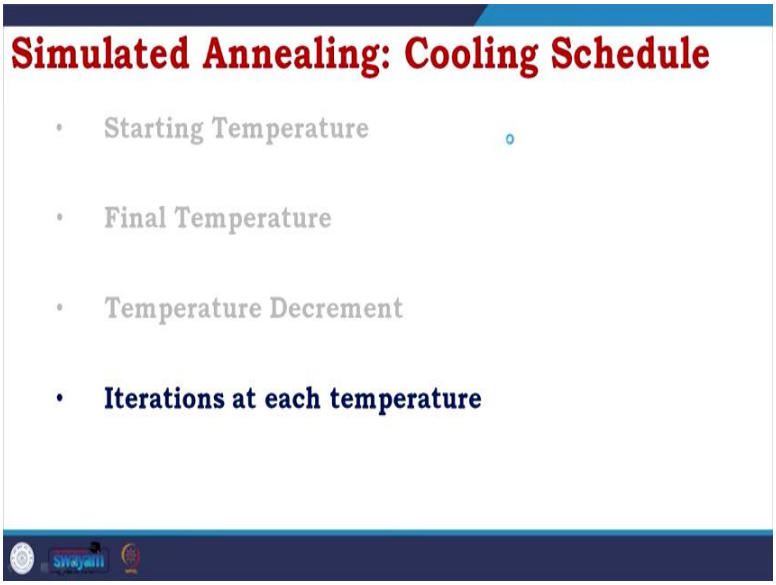
First of all, it provides thermal energy so that everything goes up to a certain energy level. Now, they are cooling down how you know they are coming to minima. First of all, you use that random number as a function, and you perform dy / dx when dy / dx = 0, you are conformed that you are in a peak. It could be maxima. It could be minima you attend a peak. Now you do double derivative, d2y / dx2 that is greater than 0 less than 0 tells you it is maxima or minima. Now, you calculate the energy suppose you get -5 kilocalories, you repeat the process, and in the next, you get -7 kilocalories. So, -7 kilocalories means you are at a better minimum than the other.

(Refer Slide Time: 01:01:38)



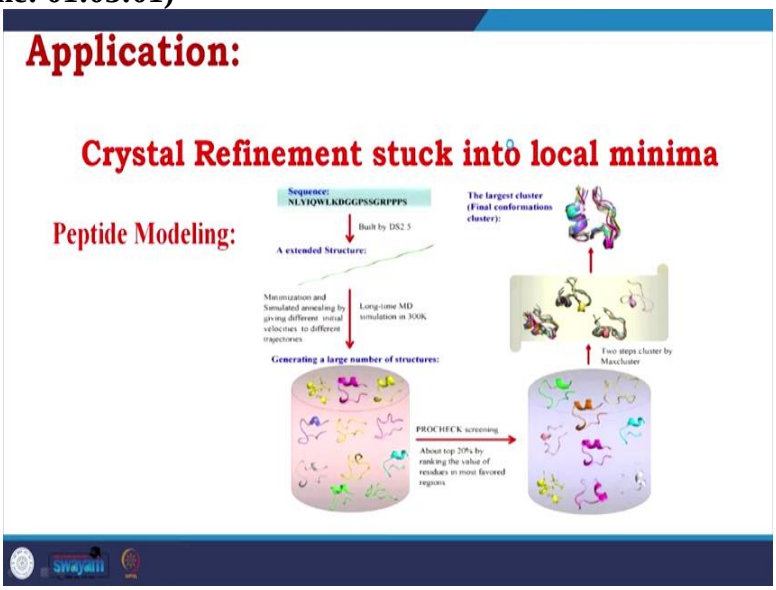
This is what I talked about in protein. In butane, when you move the conformers, you get different maximas and put maxima and minima. See, these minima is lower than this. So, if this is local, this is global. So, as I told global and local maximum, local minima, local minima, global minima, you get them, and this is the process I already talked about.

(Refer Slide Time: 01:02:34)



So, the cooling schedule will be maintained first at the starting temperature, final temperature, then temperature decrement and iterations at each temperature. The temperature choice will be made according to the requirement or setup of the experiment.

(Refer Slide Time: 01:03:01)



Crystal refinements stuck into local minima; If you remember, I try to talk details about the are factors when you perform minimization in each of your cycles, you have observed there is a minimization of the R. So, your journey starts at 0.6 to 0.7, 60% to 70%, and you have to achieve point to suppose in between after multiple refinements, you are not being able to see the reduction of the R factor, you think that you stuck into the local minima.

In that time, you have to apply simulated annealing to get out of this local minima and then continue.

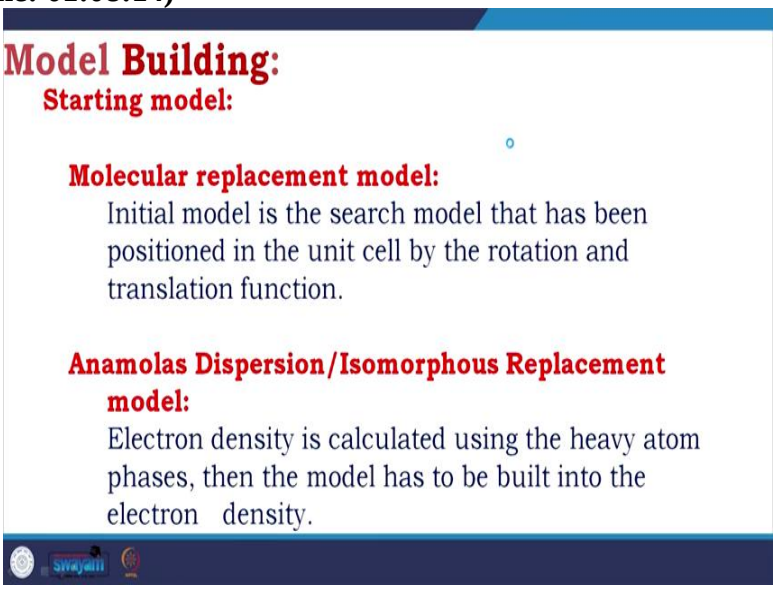
Why peptide modeling is becoming very important nowadays if you look at the therapeutic scenario before people said it is the proteins you should consume, and now the therapeutic spectrum is altering.

Now, people are saying it is not the protein rather than the peptides. Your protein will go inside the body, protein will be digested converted into a peptide, and a peptide would be going and binding to the receptors to activate certain things. What is the difference between modeling a protein and modeling a peptide?

Many people take the coordinates cut out from the protein and say that this is the model structure of the peptide. You have to take care of the increasing degrees of freedom of the peptides. So that is why here is the peptide modeling algorithm, which takes the sequence and then instead of doing long time MD simulation, minimization, and simulated annealing by giving different initial velocities to different trajectories.

Then it generates many structures, and the whole structure assemblies go through the project screening were about top 20% by ranking the value of residues in the most favored region. So, you have all possible of them, and you get the 20% now, you do the clustering, and you get the largest cluster to be chosen as the final confirmation peptide. So in that way, simulation-based peptide modeling would be used.

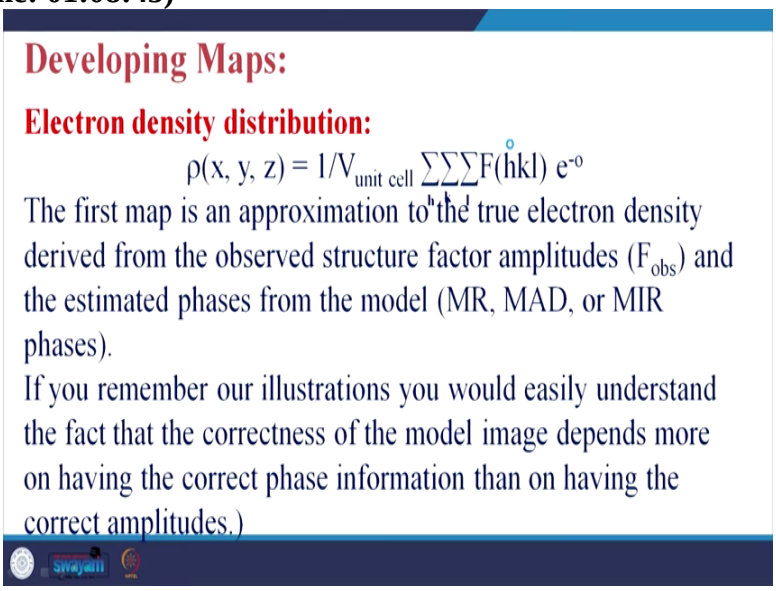
(Refer Slide Time: 01:08:14)



Molecular replacement model: The initial model is the search model that has been positioned in the unit cell by the rotation and translation as we talked about 6 degrees of freedom, three rotational degrees of freedom, and three translational degrees of freedom. Anomalous dispersing isomorphous replacement model: Which comes from the heavy atom, electron

densities calculated using the heavy atom phases, then the model has to be built into the electron density.

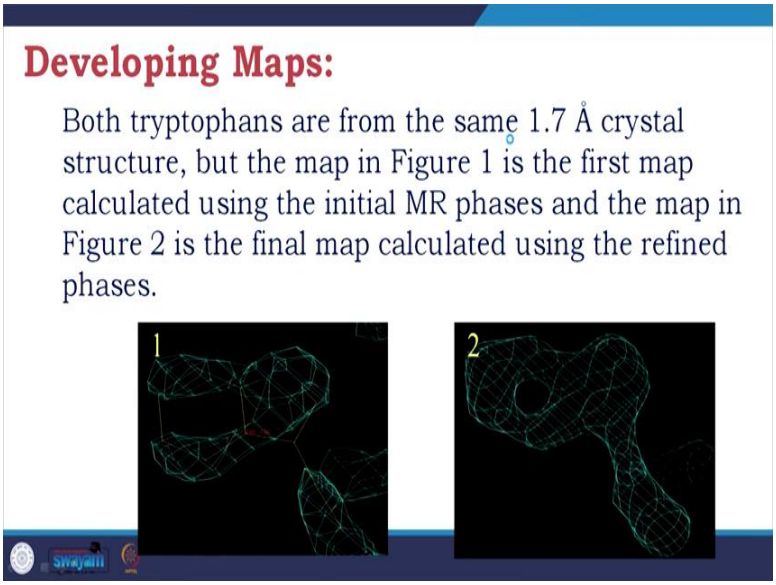
(Refer Slide Time: 01:08:45)



Now we will go for developing maps of electron density distribution. We know the formula the first map is an approximation to the true electron density derived from the observed structure factor amplitude (F_{obs}), which is observed, and the estimated phases from the model coming from molecular replacement or heavy atom were MAD, MIR, Myra, Cyras, every possible method are applied.

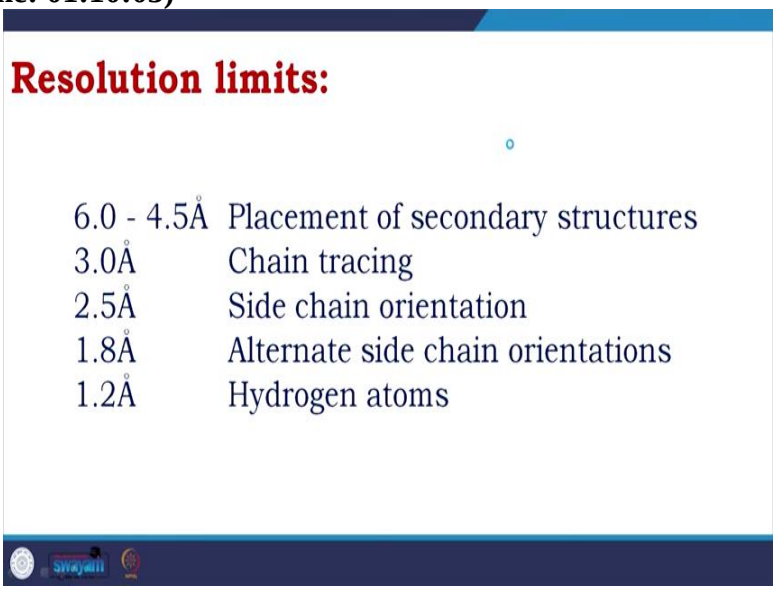
If you remember our illustrations, you would easily understand the fact that the correctness of the model image depends more on having the correct phase information than on having the correct amplitude remember the cat experiment.

(Refer Slide Time: 01:09:34)



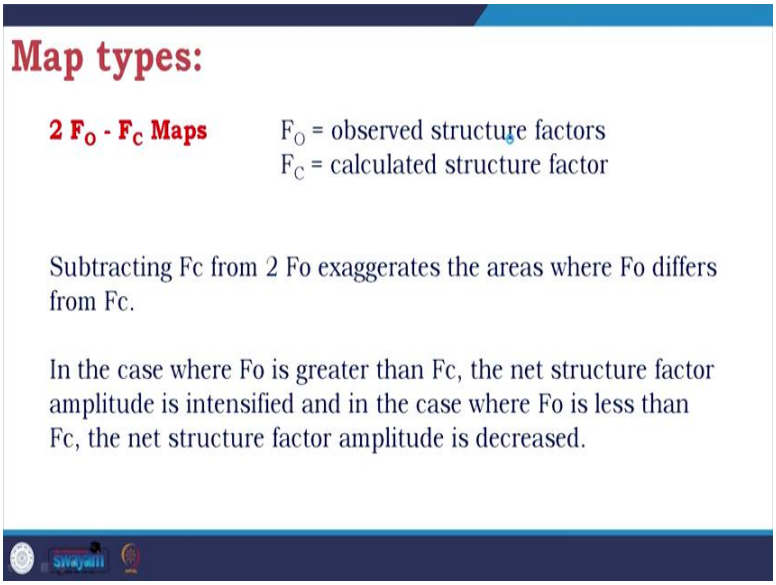
Both tryptophans are from the same 1.7-angstrom crystal structures, but the map in figure 1 is the first map calculated using the initial molecular replacement phases, and the map in figure 2 is the final map calculated using the refined phases because it is defined phases. So here, get a much better electron density map.

(Refer Slide Time: 01:10:03)



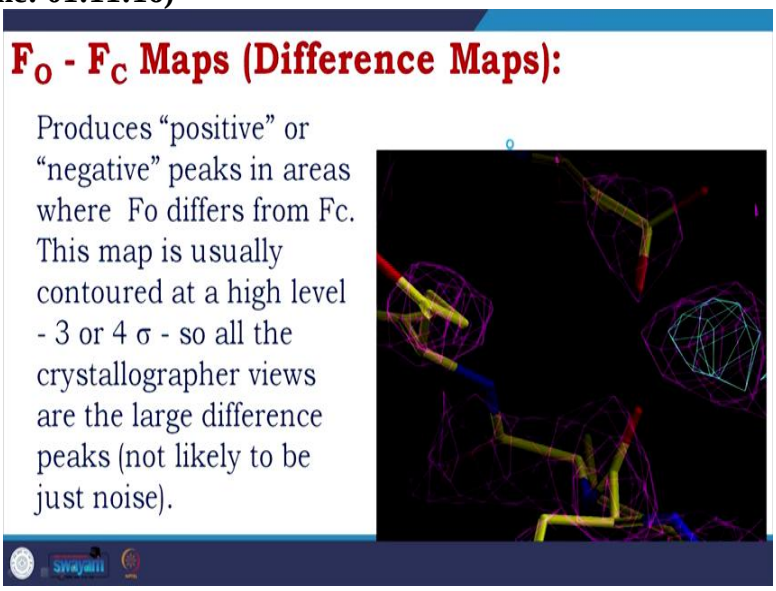
Resolution limit is very important 6 to 4.5 angstrom is the placement of the secondary structures. Three angstroms is for main-chain tracing, 2.5 angstroms for side-chain orientation, 1.8 angstroms for alternate sidechain orientations, 1.2 angstroms for looking at the effect of hydrogen atoms.

(Refer Slide Time: 01:10:29)



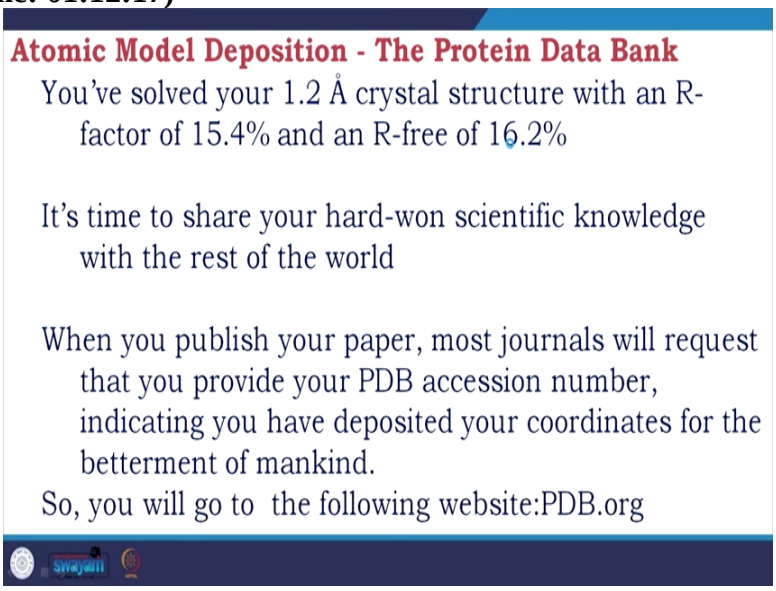
Map types: 2 F_0 – F_C map; F_0 is observed structure factor, F_C is the calculated structure factor, subtracting F_C from $2F_0$ or Fobs exaggerates the areas where F_0 differs from F_C. In the case where F_0 is greater than F_C, the net structure factor amplitude is intensified and the case where the F_0 is less than F_C, the net structure factor amplitude is decreased.

(Refer Slide Time: 01:11:16)



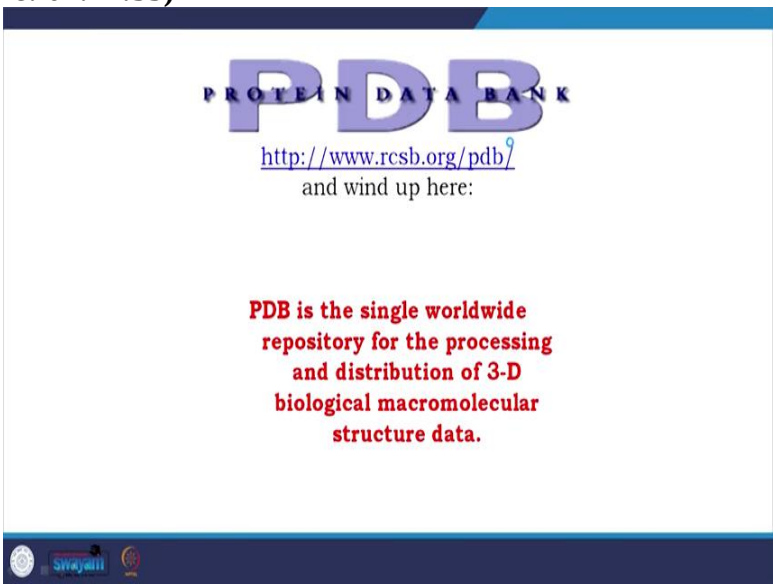
F_0 – F_C map (Difference Maps): Produces positive or negative peaks in areas where F_0 differ from F_C. This map is usually contoured at a high level, and you get the contoured level by looking at the sigma value. So, all crystallographer's views are the different large peak not likely to be just noise. So if you do it in that way, you could see the actual electron density map and not the noise. So now we have done everything. So let us deposit the model in the protein databank. So you absorbed your 1.2-angstrom crystal structure with an R-factor of 15.4 and R-free of 16.2.

(Refer Slide Time: 01:12:17)



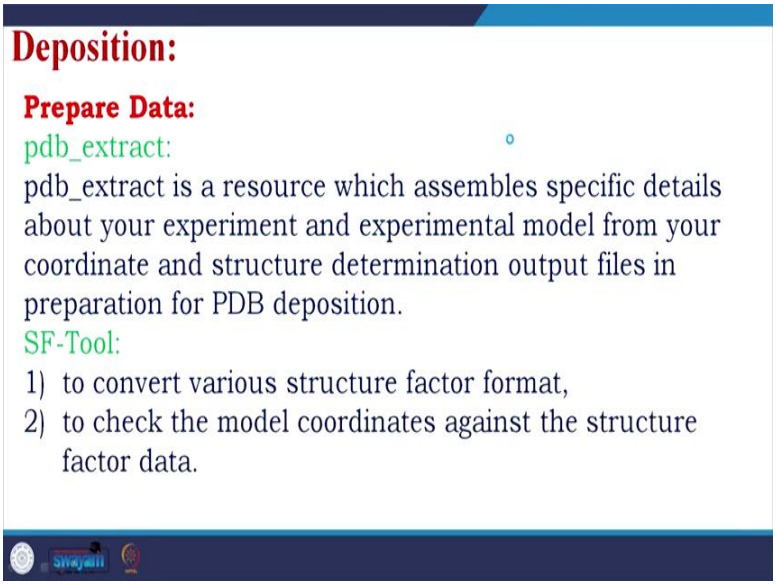
It is time to share your hard-won scientific knowledge that is the world when you publish your paper. Most journals will request that you provide your PDB accession number to add it on that nowadays it is not only the PDB accession number that there is a validation certificate coming from PDB you have to submit that only then that journal would agree to the fact that you have solved the structure. So indicates you have deposited your coordinates for the betterment of humankind. So then you go to the following website pdb.org.

(Refer Slide Time: 01:12:55)



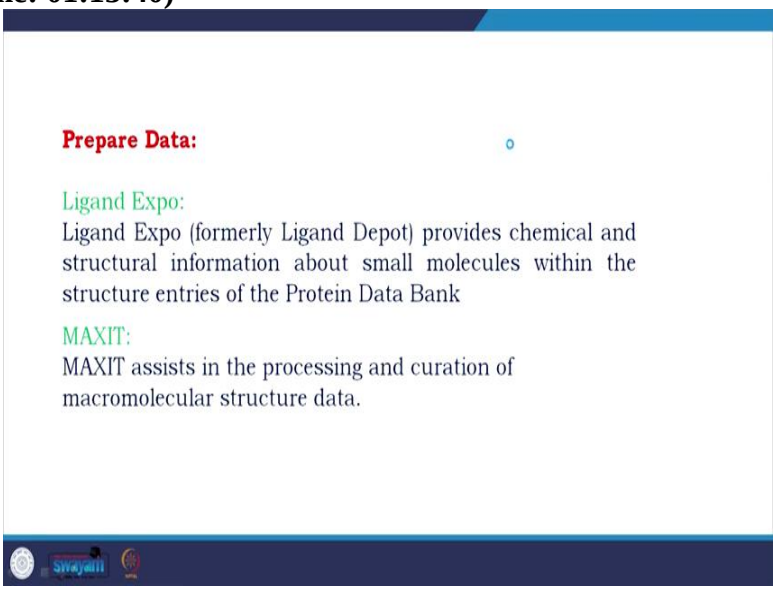
This is a protein databank, and PDB is the single worldwide repository for the processing and distributing of 3D biological macromolecules structure data.

(Refer Slide Time: 01:13:10)



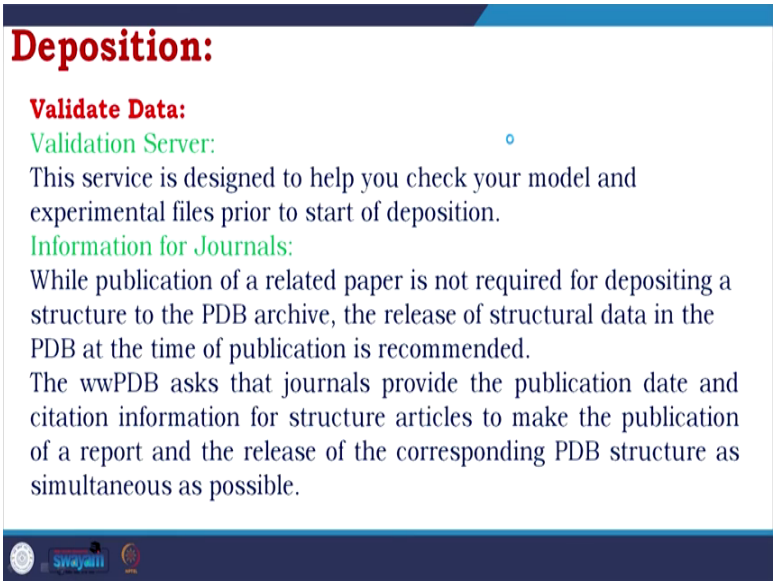
For deposition, you have to prepare data, pdb-extract, it is a resource that assembles specific details about your experiment and experimental model from your coordinate and structure determination output files in preparation for PDB deposition. SF-tool to convert various structure factor formats and check the model coordinates against the structure factor data.

(Refer Slide Time: 01:13:40)



Ligand Expo (formerly ligand depot) provides chemical and structural information about small molecules within the structure entries of the protein databank. MAXIT assists in the processing and curation of macromolecules structure data.

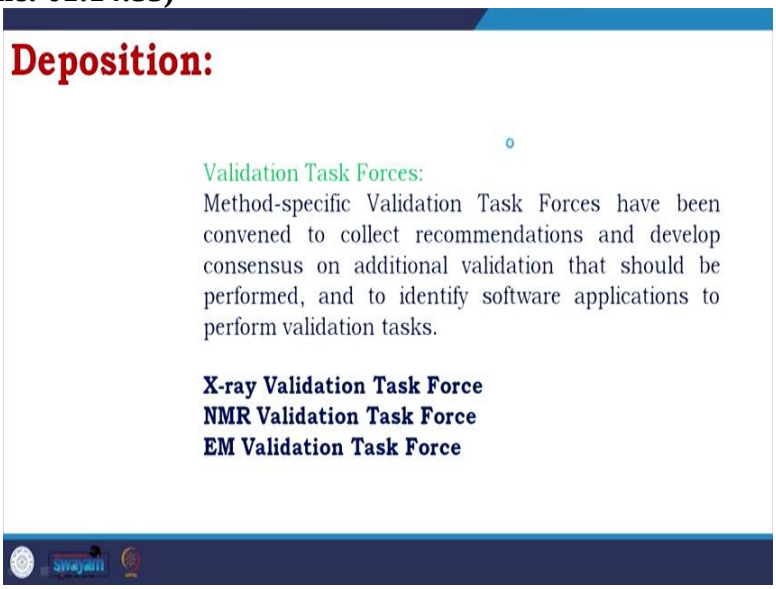
(Refer Slide Time: 01:13:53)



You also have to do validation, and there are validation servers. This service is designed to help you check your model and experimental files before start-up deposition information for journals. At the same time, publication of a related paper is not required for depositing a structure to the PDB archive. The release of structural data in the PDB at publication is recommended.

The wwPDB asks that journals provide the publication date and citation information for structure articles to make publication of a report and release the corresponding PDB structure as simultaneous as possible. So you deposit it is there, then when you send the related work to the journal, when the journal publishes the paper, PDB will release your structure.

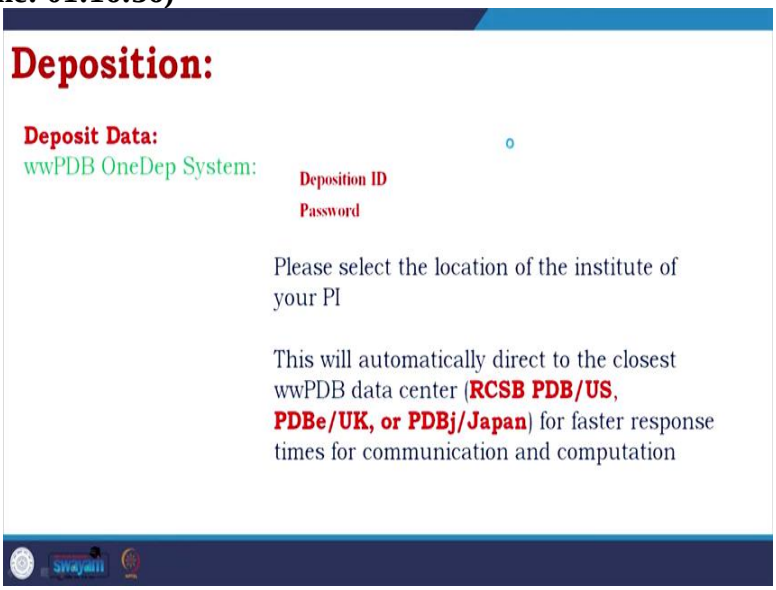
(Refer Slide Time: 01:14:55)



Validation task forces methods specific validation task forces have been convened to collect recommendations and developed consensus on additional validation that should be performed and identify software applications to perform validation tasks.

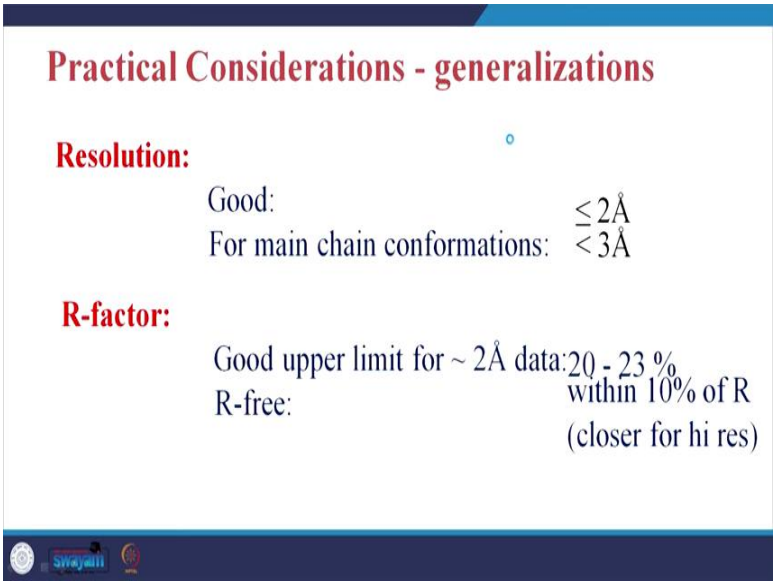
So, there are X-ray validation task force, NMR validation task force, electron microscope validation task force. They are like a crystallographer, NMR specialists, look at other factors which the automation programs are not being able to detect, and they will send you again to improve those factors.

(Refer Slide Time: 01:16:36)



So, deposited, deposit data wwPDP oneDEP system you have deposited ID, your password and you select the location of the Institute of your PI, or if you are the PI, this will automatically direct to the closest wwPDB data center RCSB PDB is for the US, PDBe or UK and PDBj for Japan. So, from where you are closest to, you could also want to submit to a a certain place for faster response times for communication and computation.

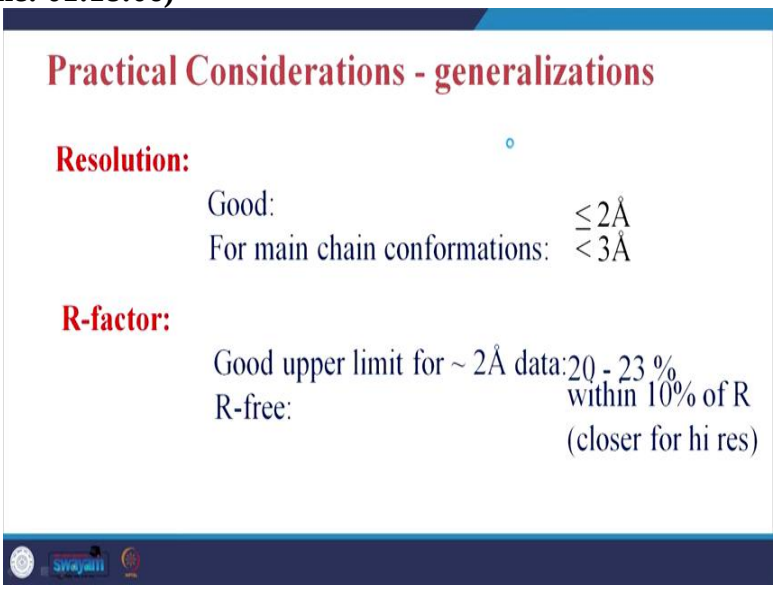
(Refer Slide Time: 01:17:22)



Validation suite there are procheck, nucheck sfcheck. Procheck I will discuss assessing the geometry of the residues in a given protein structure compared with stereochemical parameters derived from well-refined high-resolution structures. Unusual regions highlighted by procheck are not necessarily errors but maybe unusual features for a reasonable explanation.

For example, distortion due to ligand-binding the protein's active site; nevertheless, regions should be checked carefully.

(Refer Slide Time: 01:18:06)



Now, practical considerations, for resolution like less than 2 angstroms, for main-chain conformation less than 3 angstroms. R-factor: good upper limit for 2-angstrom data 20 to 23%, R-free would be within 10% of the R work and closer for the high-resolution data.

(Refer Slide Time: 01:18:34)