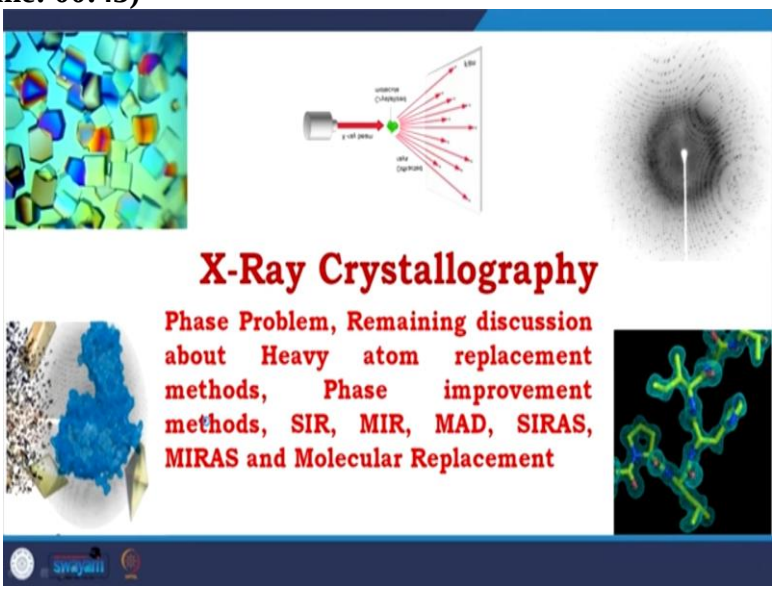
Structural Biology
Prof. Saugata Hazra
Department of Biotechnology
Indian Institute of Technology, Roorkee

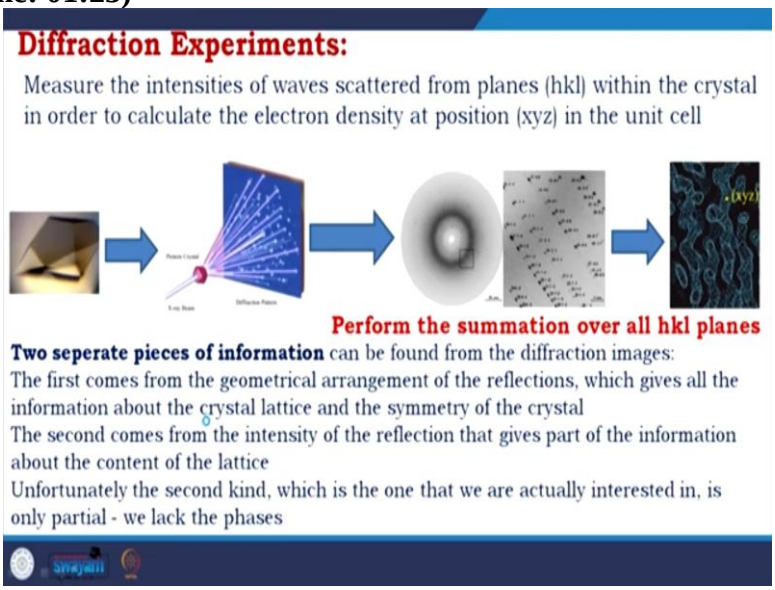
Lecture - 24
X-Ray Crystallography - Phase Problem - Part 2

(Refer Slide Time: 00:43)



We are actually at the end of our course document on crystallography. We discussed phase problems in the last class. Today we will again discuss a few parts of the phase problem, the remaining discussions about the heavy atom replacement methods, phase improvement technologies, SIR, MIR, MAD and a little introduction to SIRAS, MIRAS and then we will talk about another technique which is the molecular replacement to solve the phase problem.

(Refer Slide Time: 01:23)

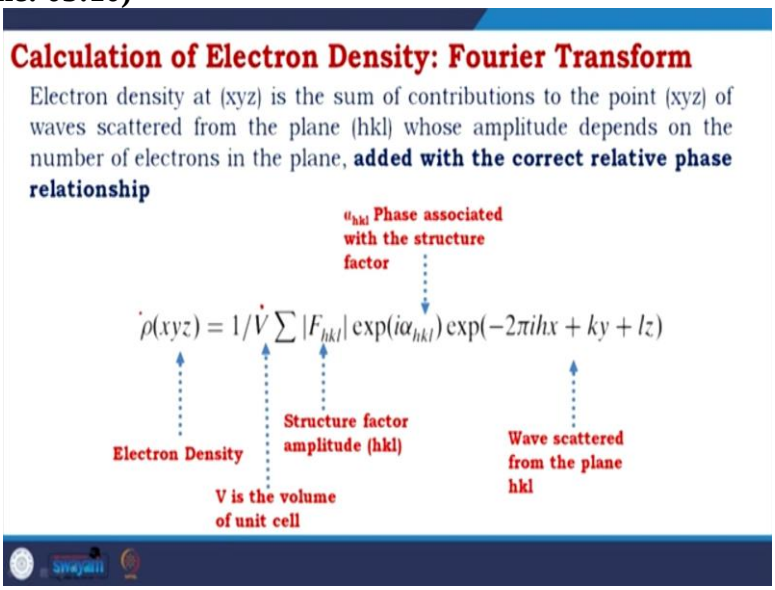


So, if I go back a bit and start where we discussed a lot, we clone a protein, purify a protein and get the crystal. So, we get the crystal, and we do a diffraction experiment. We measure the intensities of waves scattered from planes within the crystal to calculate the electron density at position xyz in the unit cell.

So, we perform the summation over all hkl planes. What we get from our diffraction experiment is two separate pieces of information. The diffraction image gives us the first information, which is the geometrical arrangement of the reflection gives all the information about the crystal lattice and the symmetry of the crystal. The second information we get from the intensity of the reflection gives part of the information about the content of the lattice.

And as I was discussing throughout the last few lectures in the second one, the intensity information gives us the amplitude, but we are not getting the phase; hence, it is the phase problem.

(Refer Slide Time: 03:10)

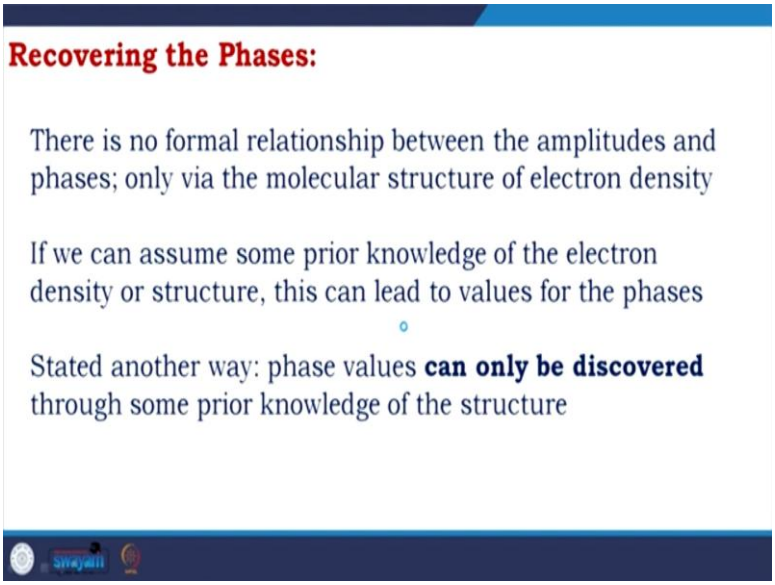


So, we have to do the calculation of the electron density and we have to perform Fourier transform. So, electron density at xyz is the sum of contribution to the point xyz of waves scattered from the plane hkl whose amplitude depends on the number of electrons in the plane added with the correct relative phase relationship. So, if we again look at the famous formula which we are continually looking at

$\rho(xyz) = 1/V \sum \left| F_{hkl} \right| \exp(i\alpha_{hkl}) \exp(-2\pi i hx + ky + lz)$

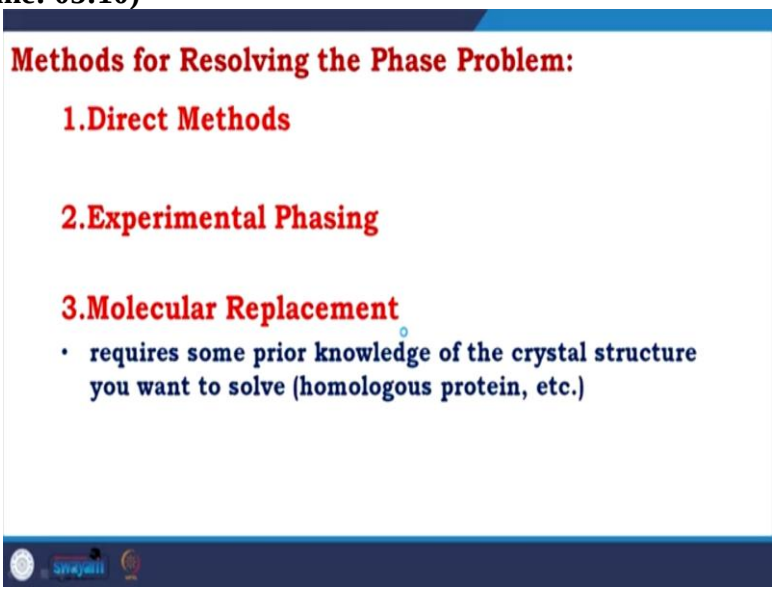
$\rho(xyz)$ is the electron density, V is the volume of the unit cell, the absolute value of $\left| F_{hkl} \right|$ is the structure factor amplitude, α_{hkl} phase is associated with the structure factor, and the waves scattered from the plain hkl.

(Refer Slide Time: 04:13)



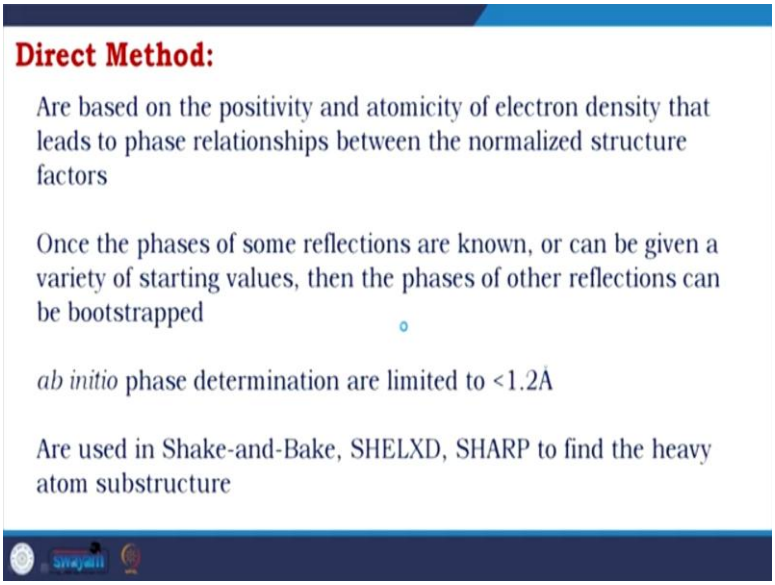
So, as we are continuously talking about how to get the phase, the phase problem, and how to recover the phase, there is no formal relationship between the amplitude and the phases but only via the molecular structure of electron density. If we can assume some prior knowledge of the electron density or structure, this can lead to values of the phases. So, phase value can only be discovered through prior structure knowledge.

(Refer Slide Time: 05:10)



The direct method we discussed details, the experimental phasing using heavy metals we again discussed and molecular replacement which is our topic today. But, we will also talk about some very interesting features related to experimental phasing, especially mathematical ones.

(Refer Slide Time: 05:34)

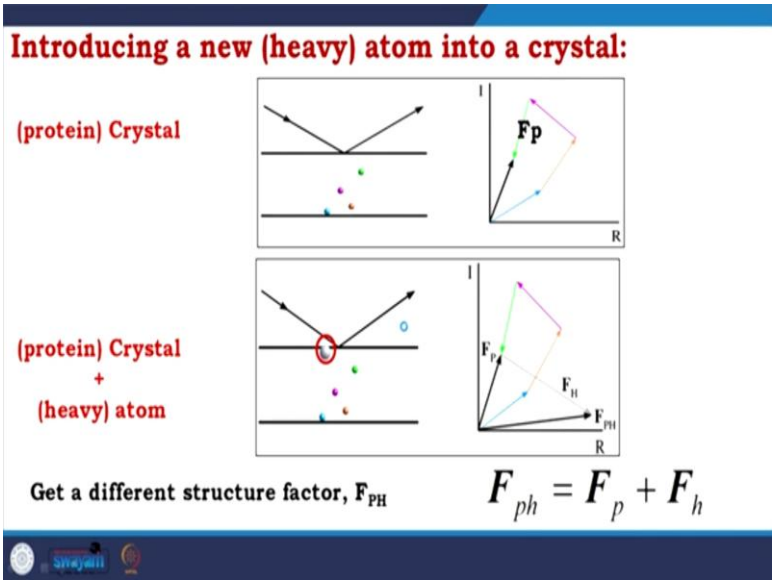


Direct method are based on positivity and atomicity of electron density that leads to phase relationships between the normalized structure factor. So, they try to extract data from the intensity, which gives the amplitude the structure factor but from that intensity that diffraction pattern, they want to get more information once the phases of some reflections are known.

Which are some reflections, some reflections where the intensities are very high or can be given a variety of starting values, then the phases of other reflections can be bootstrapped. But this ab initio phase determination is limited to less than 1.2 Angstrom resolution. So, lowering the number means you know that the higher the resolution.

But you also remember that the direct method would also help you in protein structure determination. So, they are used in Shake and Bake, SHELXD, SHARP to find the heavy atom substructure. As I told, when the intensity is very high, it is possible to correlate. So, for heavy atoms, we first know the atom and the chemistry, which are very different from the protein (protein contains carbon, nitrogen, and oxygen).

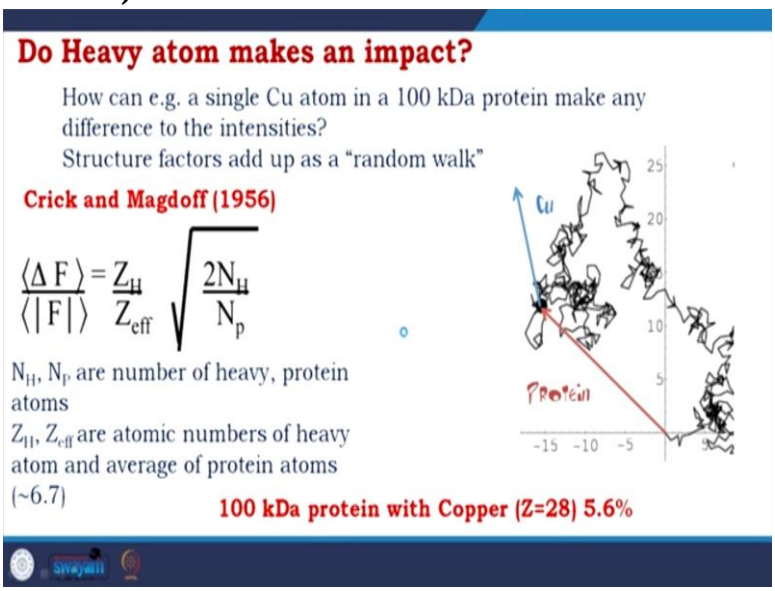
(Refer Slide Time: 07:17)



So, we could have got the identity of their information. The protein, we now introduce a heavy atom. So to calculate the Fp, the structure factor of the protein, here with the heavy atom, we get Fph, and we know that

Fph = Fp + Fh

(Refer Slide Time: 08:07)



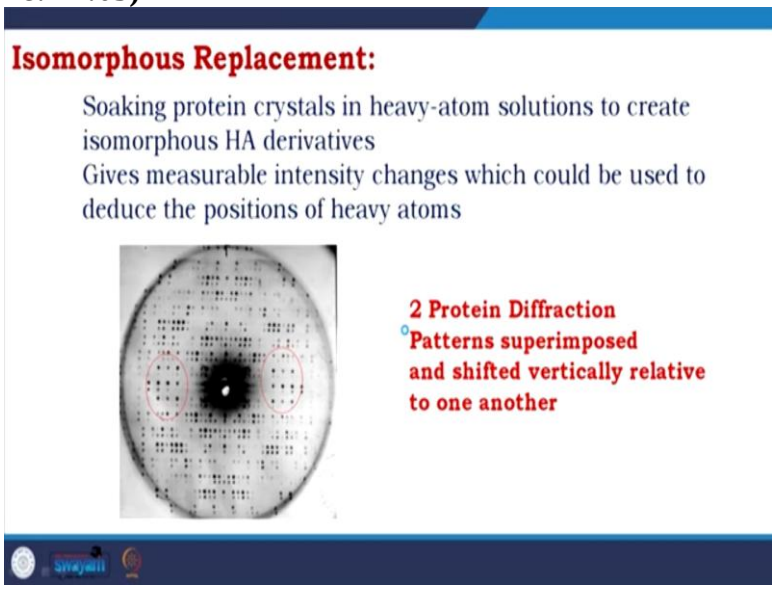
So, before going into details, one thing you should know. Does heavy atom make an impact? So, let us look at how a single copper atom in 100 KDa protein can make any difference to the intensities. So, structure factors add up as a random walk. Crick and Magdoff in 1956 have calculated this,

$\langle \Delta F \rangle / \langle | F | \rangle = Z_H / Z_{eff} \sqrt{2} N_H / Np$

N_H is the number of the heavy atom, and Np is the number of protein atoms. Z_H is the atomic number of the heavy atom, and Z_{eff} is the average of the protein atom, which is 6.7. This is the change the protein has its vector, and then the copper adds with the structure factor. So, in 100 kDa protein, copper with atomic number 28 adds 5.6%. That is the impact. Now, depending on this, you could choose the heavy metal, understand the concentration, and optimize the concentration of the heavy metal.

Why did you need to optimize this? A protein is largely made up of small hydrogen bonds. When a big metal with its high positive charge and high polarizability suddenly comes inside the protein, it is like a bombardment on the weak hydrogen bond network of the protein. A lot of time, they are deleterious. They break the crystal, and sometimes the protein starts changing shapes, and all those problems are there. Also, that is very much concentration dependent.

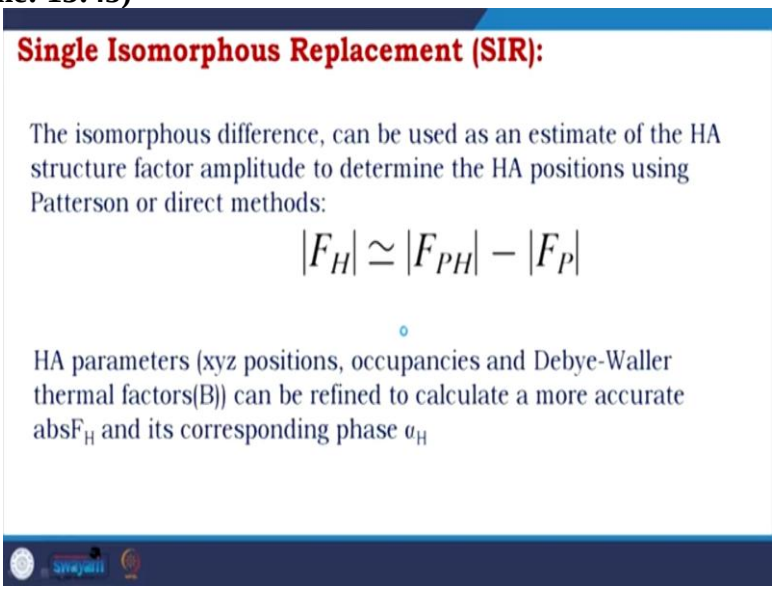
(Refer Slide Time: 12:03)



We do the soaking up protein, where we have a metal solution, we put the crystal there, or we do co-crystallization, where we add precipitants and the metal at the time of crystallization. Soaking protein crystals in heavy atoms solution to create isomorphous heavy atom derivatives gives measurable intensity changes, which could be used to deduce the position of the heavy atoms.

Two protein diffraction patterns. One is the native protein, and another is the mercury-added protein. Superimposed, we could see the shift of the peaks particularly related to one another, and we could also see the change of intensity very easily. So, this tells that by adding heavy metal there are changes.

(Refer Slide Time: 13:45)

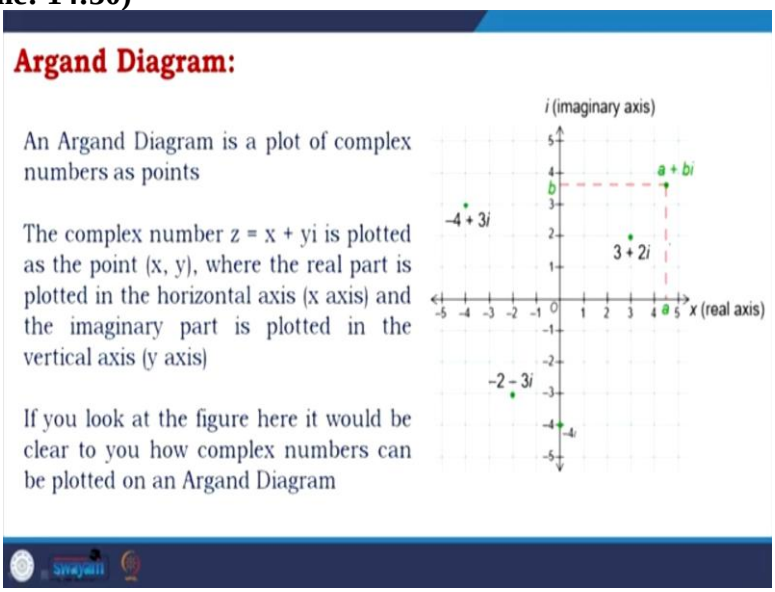


Using Patterson or direct methods, the isomorphous difference can be used to estimate the heavy atoms structure factor amplitude to determine the heavy atom position.

$|F_H| = |F_{PH}| - |F_P|$

Heavy atom parameters, such as xyz positions, occupancies and Debye-Waller thermal factor (B factor), can be defined to calculate a more accurate, absolute FH the heavy atom structure factor and its corresponding phase α_H.

(Refer Slide Time: 14:30)

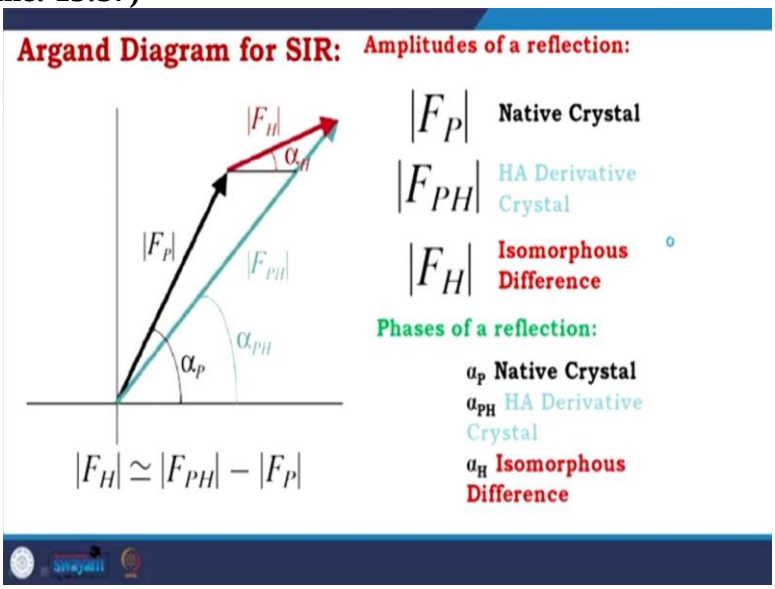


We are going to put an argument diagram and show how it is important in structure solution. The argand diagram in mathematics is a plot of complex numbers as points. The complex number, let us say

z = x + yi

is plotted as the point x y, where the real part is plotted on the horizontal axis, the x-axis, and the imaginary part is plotted on the vertical axis, the y-axis.

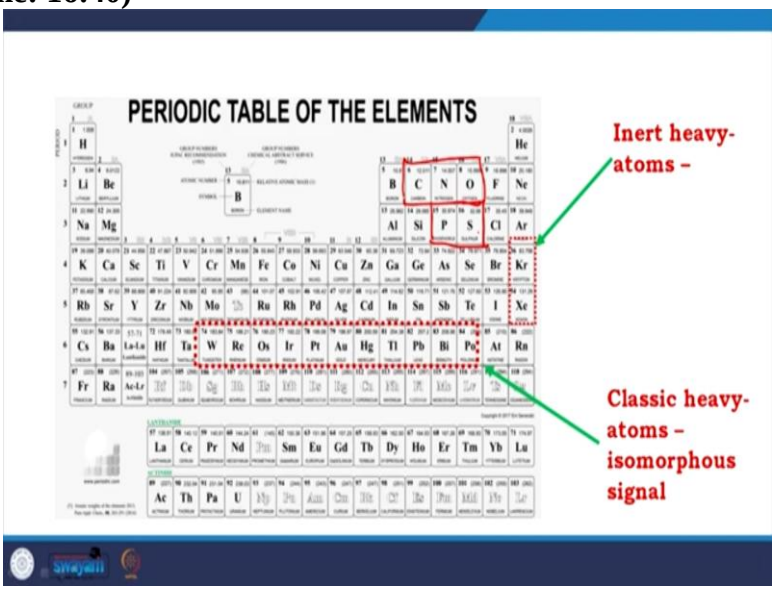
(Refer Slide Time: 15:37)



Now come to the Argand diagram of single isomorphous replacement. You will see many factors here, the amplitude of reflection $|F_P|$, which is the structure factor of the protein, $|F_{PH}|$, which is the absolute value of the structure factor of heavy metal-derived protein. And $|F_H|$ is the difference which is heavy metals contribution.

Similarly, the phases, α_P comes from the native crystal, α_{PH} heavy atom derivative crystal, and α_H isomorphous difference.

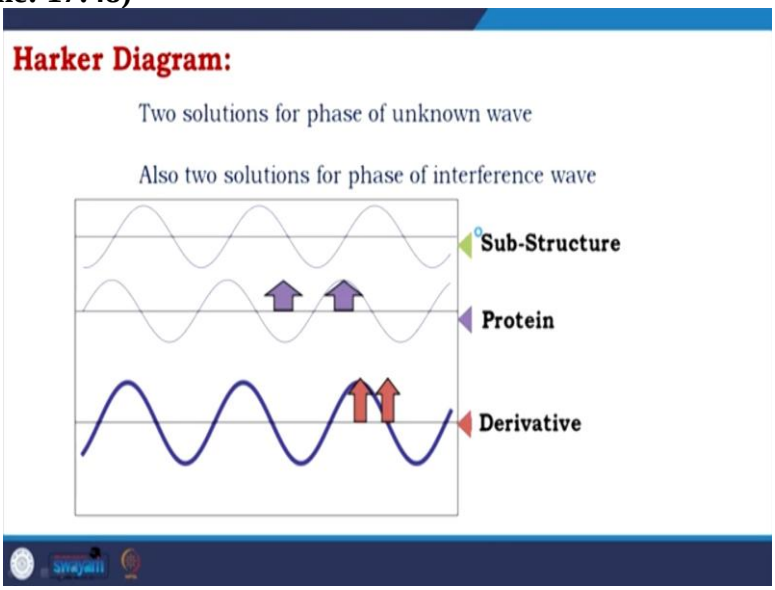
(Refer Slide Time: 16:40)



So, suppose you see the periodic table. In that case, the heavy atoms which you could choose are all in this row, which are tungsten, rhenium, osmium, iridium, platinum, gold, mercury,

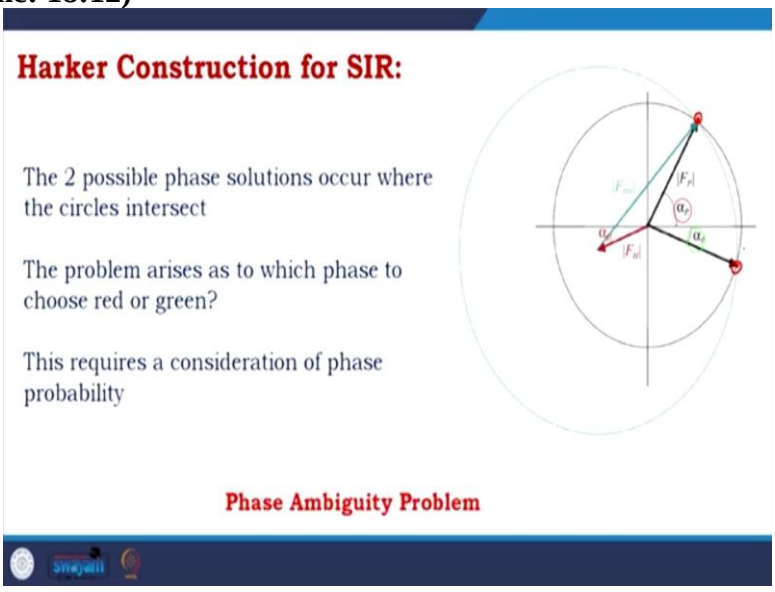
thalium, lead, bismuth, polonium. In addition, we could also use krypton and xenon which are not heavy atoms. So, you can easily see that these are much higher in electron number than carbon, nitrogen, and oxygen, which are present in normal proteins. I also add phosphorus and sulfur because sulfur is present in cysteine, methionine, and phosphorus in phosphor-modified proteins.

(Refer Slide Time: 17:48)



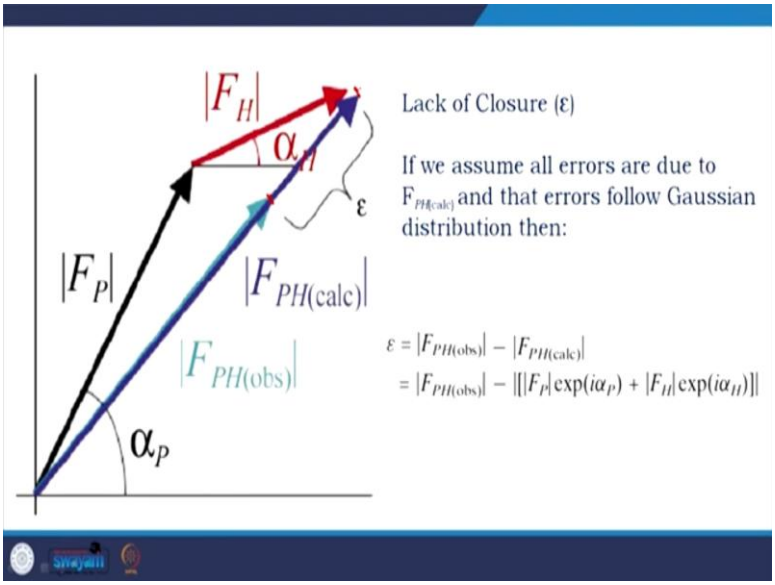
Coming to another geometrical representation called the Harker diagram, the 2 solutions for the phase of an unknown wave where meet.

(Refer Slide Time: 18:12)



Harker construction for a single isomorphous replacement that 2 possible phase solutions occur where the circle intersects. This requires a consideration of phase probability, called the phase ambiguity problem in crystallography.

(Refer Slide Time: 18:54)

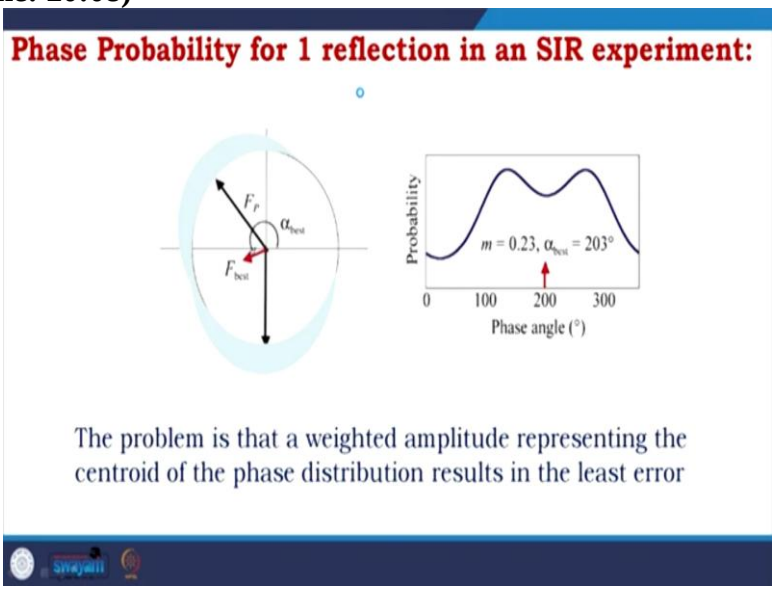


Now, we consider ε as the lack of closure, then the lack of closure

ε = equal to $|F_{PH(obs)}| - |F_{PHcalc}|$,

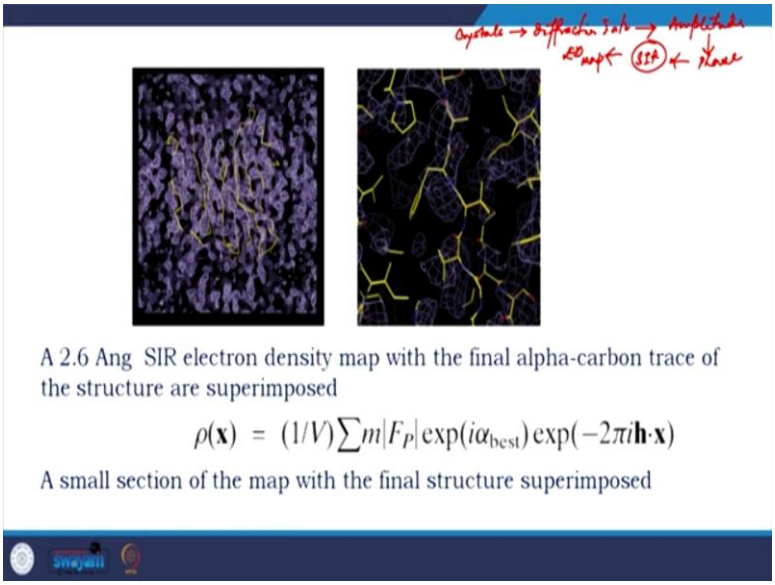
$= |F_{PH(obs)}| - [|F_P| \exp(i\alpha_p) + |F_H| \exp(i\alpha_H)]$

(Refer Slide Time: 20:08)



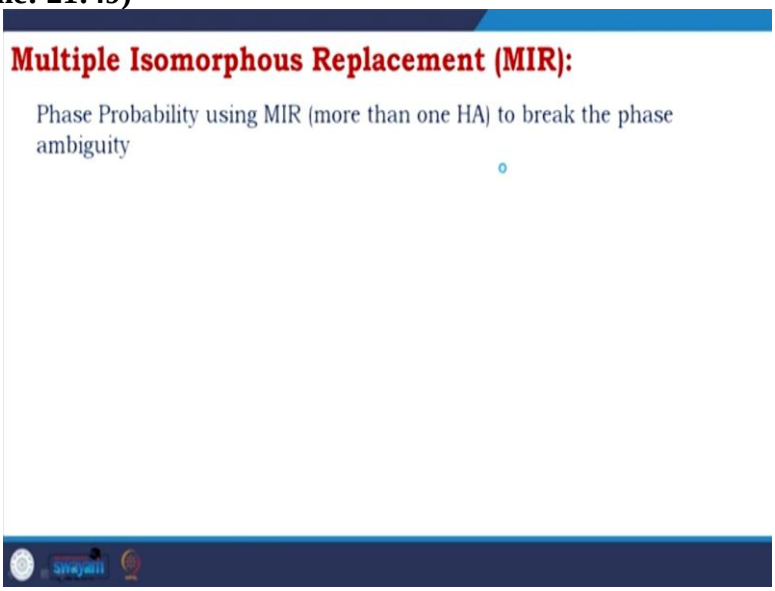
The phase probability for one reflection in an SIR experiment, the problem is that a weighted amplitude representing the centroid of the phase distribution result in the least error.

(Refer Slide Time: 20:30)



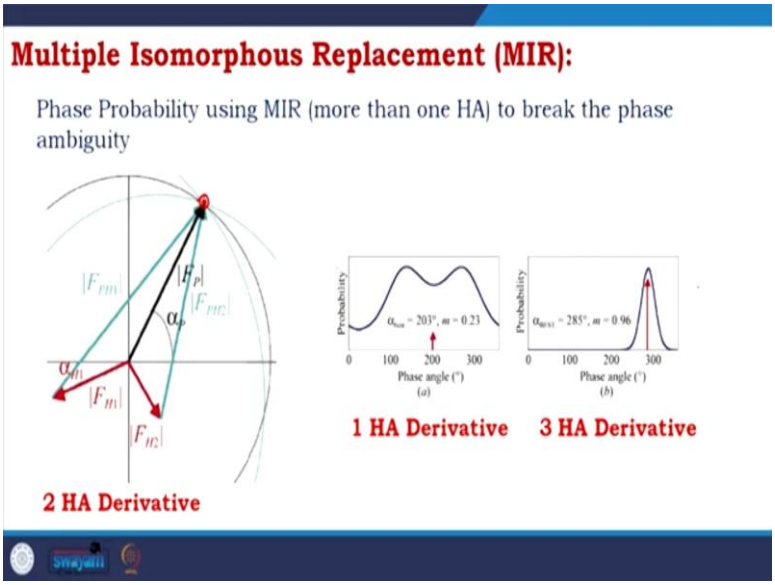
A 2.6 angstrom single isomorphous replacement electron density map with the final alpha carbon trace of the structure superimposed. So, we had crystals, we got the diffraction data, we tried to solve, it we got amplitude but we did not get the phase. Now we are putting single isomorphous replacement and we get the ED map, this is a 2.6 angstrom resolution density map. But and this is a small part of that where the final structure is superimposed. So, here you see the alpha carbon trace, here in a zoomed one, you see the side chains included here.

(Refer Slide Time: 21:49)



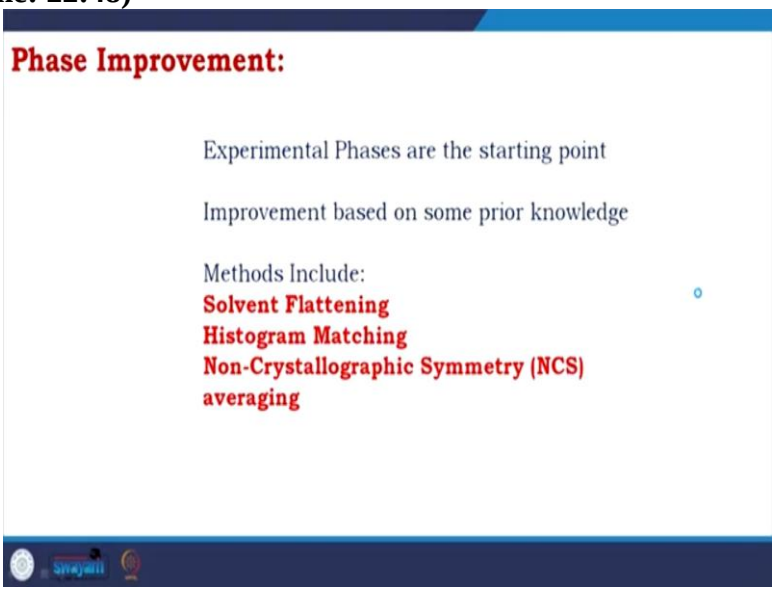
Multiple isomorphous replacement, in SIR we are using one heavy atom derivative, here we are actually using more than one heavy atom to break the phase ambiguity.

(Refer Slide Time: 22:11)



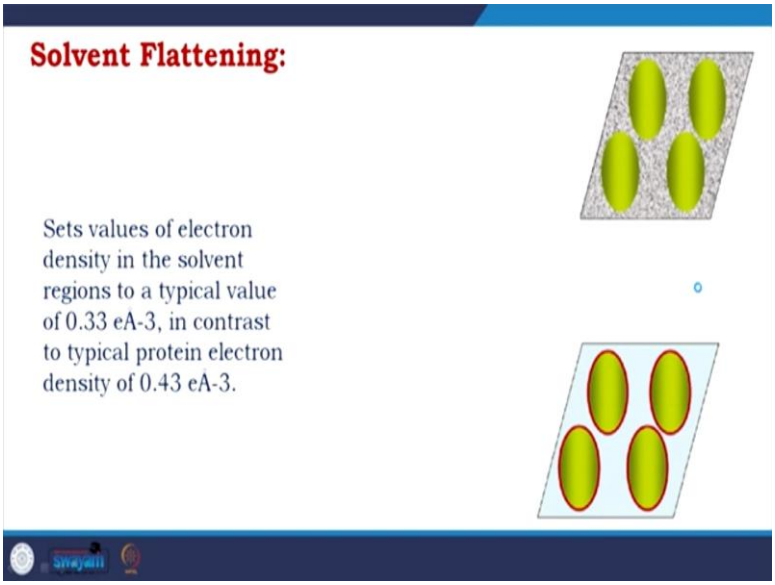
Instead of one, we use two, there is no ambiguity. If you see that, if we compare two heavy atom derivative, this is a one heavy item derivative, you get that confused two peak, and the 3 heavy atom (3 HA) derivative you get a very sharp peak. So, the problem of phase ambiguity is solved.

(Refer Slide Time: 22:48)



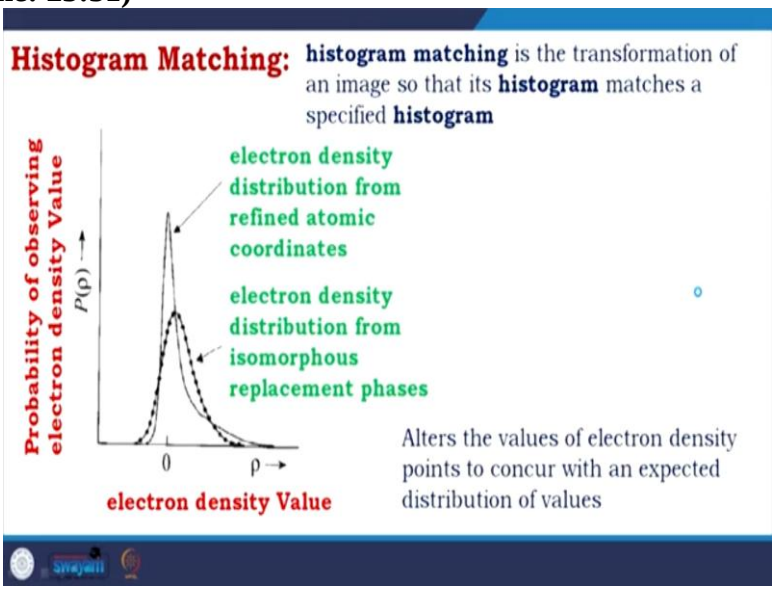
We could improve the phase: Experimental phases are the staring point. So, we did the heavy atom replacement and get the experimental phase. The improvement based on some prior knowledge. The methods are solvent flattening, Histogram matching and non crystallographic symmetry(NCS), and averaging.

(Refer Slide Time: 23:31)



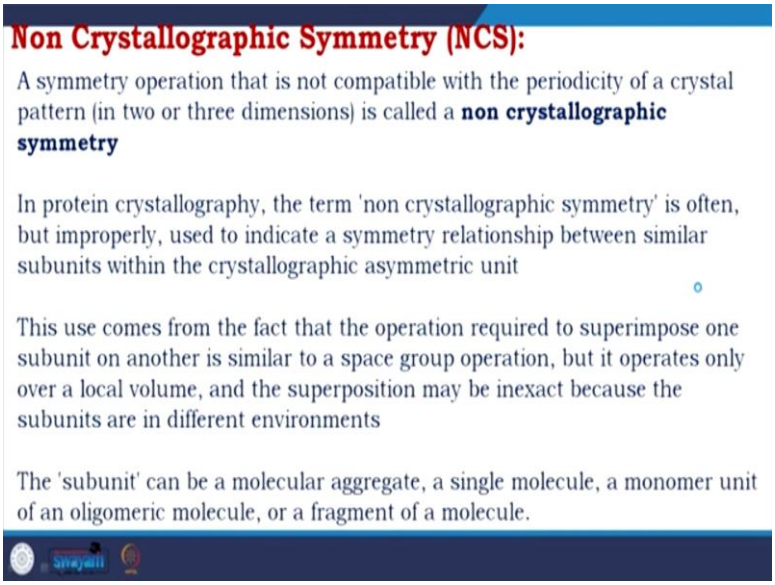
Sets the values of electron density in the solvent region to a typical value which is given here in contrast to the typical protein electron density. So actually we are differentiating between a protein and a solvent, so that the solvent could develop its phase better.

(Refer Slide Time: 23:31)



Histogram matching is the transformation of an image, so that its histogram matches a specified histogram. So, if you look at, we are talking about two peaks, which are plotted against electron density value (in the x axis) and probability of observing electron density value. We get two peaks, electron density distribution from refined atomic coordinates and electron density distribution from isomorphous replacement phases. It alters the values of electron density points to concur with an expected distribution of values.

(Refer Slide Time: 25:35)

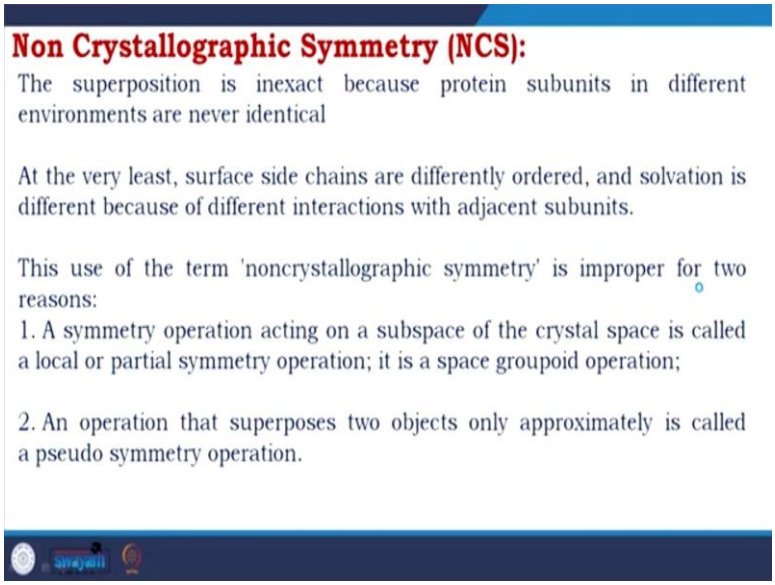


A symmetry operation that is not compatible with the periodicity of a crystal pattern (in 2 or 3 dimension) is called a non crystallographic symmetry.

In protein crystallography, the term non crystallographic symmetries is often, but importantly used to indicate a symmetry relationship between similar sub units within that crystallographic asymmetric unit.

This use comes from the fact that, the operation required to super impose one sub unit on another is similar to a space group operation, but it operates only over a local volume. And the superposition may be inexact because the sub units are in different environments. The subunit can be a molecular aggregate, a single molecule, a monomer unit of an oligomeric molecule or a fragment of a molecule.

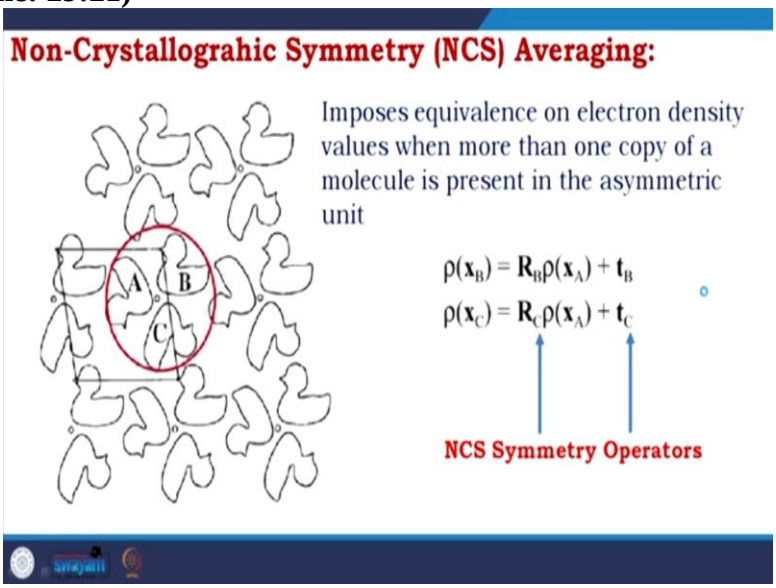
(Refer Slide Time: 27:38)



The superposition is inexact because proteins subunit in different environments are never identical. At the very least, surface side chains are differently ordered and solvation is different because of different interaction with adjacent subunits.

The use of the term non crystallographic symmetry is improper for two reasons. One is, a symmetry operation acting on a subspace of a crystal space is called local or partial symmetry operation, it is a space groupoid operation. Second, an operation that superposes 2 objects only approximately is called a pseudo symmetry operation.

(Refer Slide Time: 29:21)

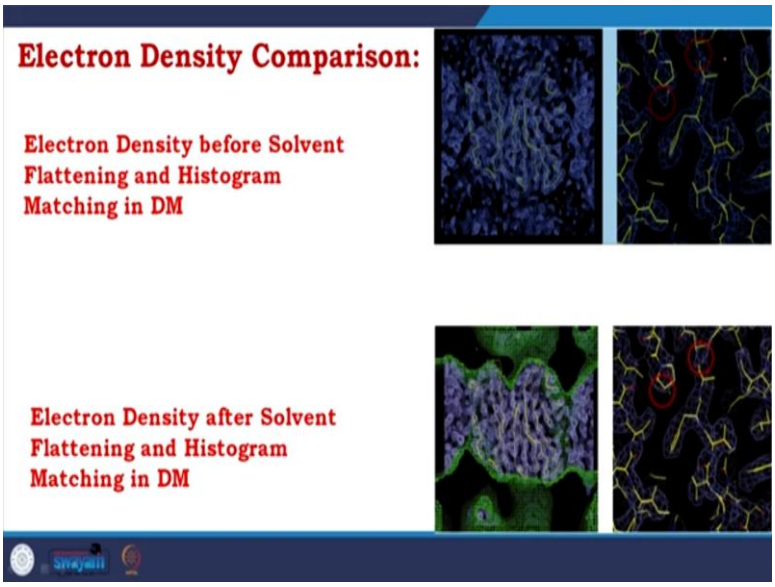


You see A, B and C and they imposes equivalence on electron density values when more than one copy of a molecule is present in the symmetric unit.

$\rho(X_B) = R_B\rho(X_A) + t_B$

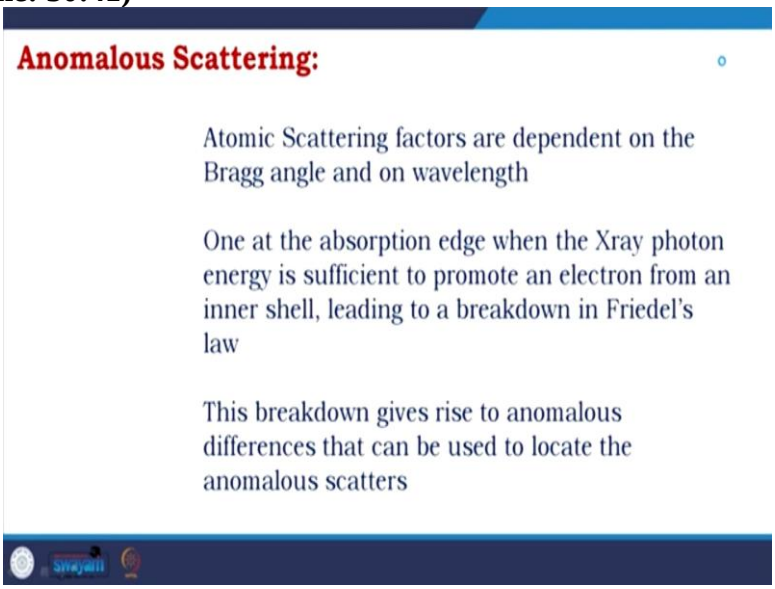
$\rho(X_C) = R_C\rho(X_A) + t_C$

(Refer Slide Time: 29:51)



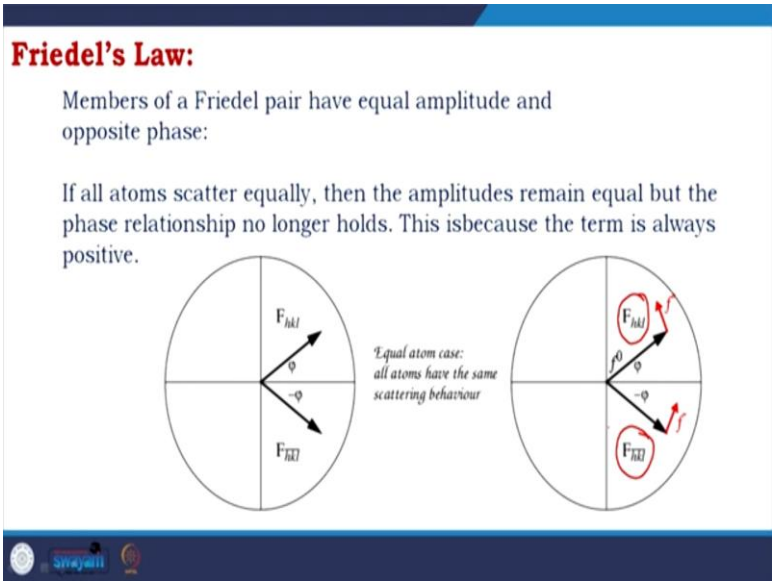
We have done all those solvent flattening, histogram matching and all you see there is a beautiful change, the maps are sharper the phase is definitely improved and that is where the critical role of these methods solvent flattening, histogram matching and averaging of non crystallographic symmetry.

(Refer Slide Time: 30:41)



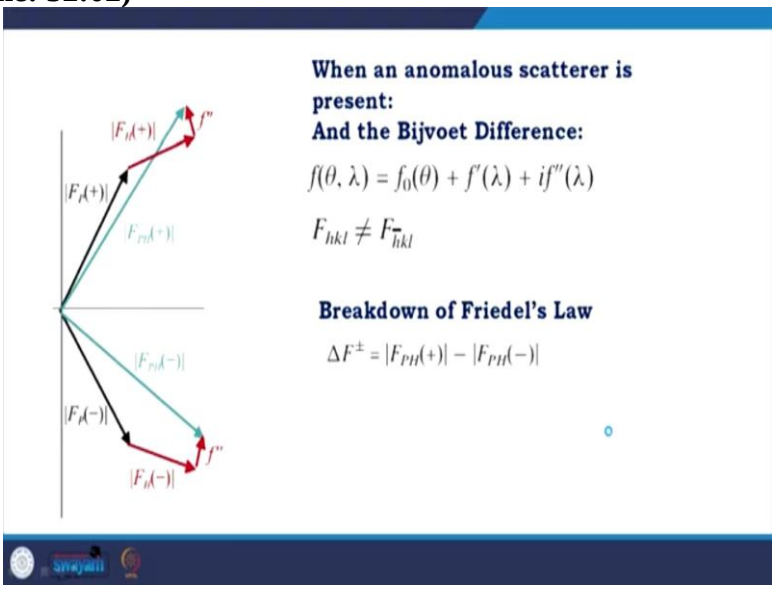
Atomic scattering factors are dependent on the Bragg angle and on wavelength. One at the absorption edge when the X ray photon energy is sufficient to promote an electron from an inner cell, leading to a break down in Friedels law. This breakdown gives rise to anomalous differences that can be used to locate the anomalous scatterers.

(Refer Slide Time: 31:17)



The members of a Friedel pair have equal amplitude and opposite phase. If all atoms scatter equally then the amplitudes remain equal but the phase relationship no longer holds. This is because the term is always positive. So, if you compare an equal atom case, all atoms have the same scattering behavior. So, when you compare between F_{hkl} and F_{hkl}^{-} you would not get, the difference.

(Refer Slide Time: 32:02)



When an anomalous scatter is present there is by void differences,

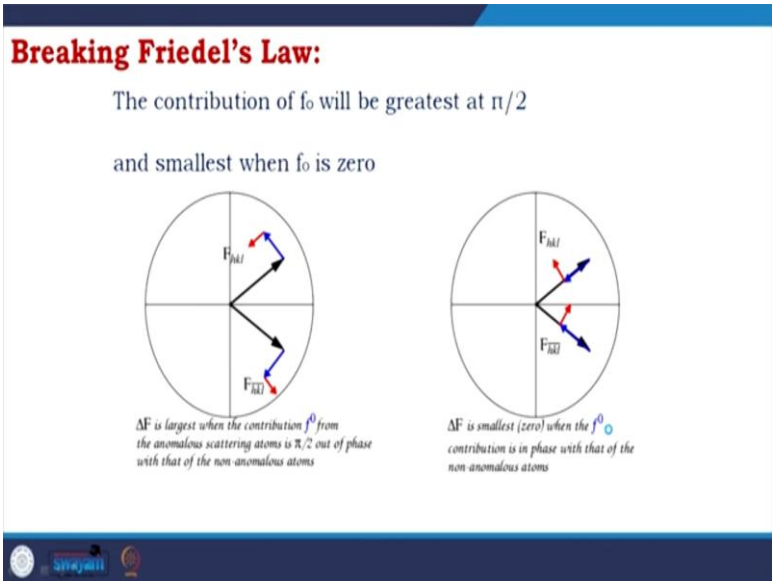
$f(\theta,\lambda) = f_0(\theta) + f'(\lambda) + if''(\lambda)$

$F_{hkl} \neq F_{hkl}^{-}$

And the breakdown of Friedels law is

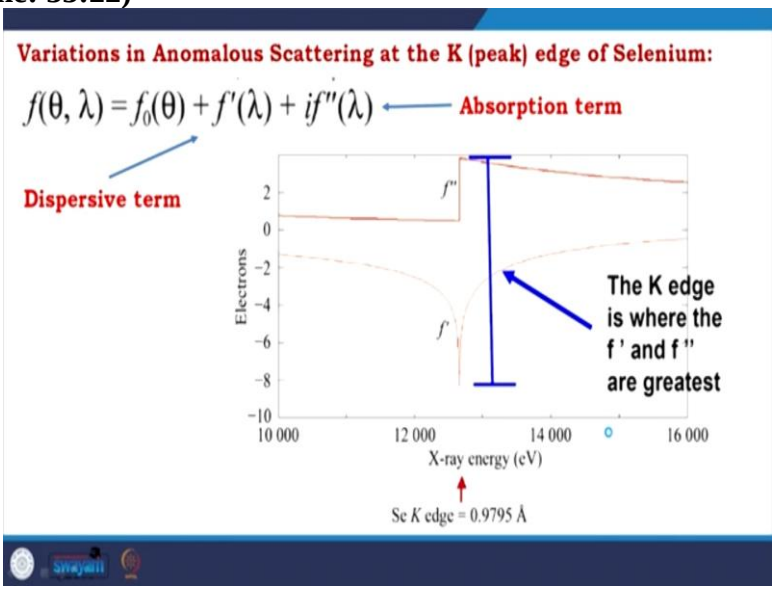
$\Delta F^{\pm} = |F_{PH}(+)| - |F_{PH}(-)|$

(Refer Slide Time: 32:39)



The contribution of f_0 will be greatest at $\pi / 2$ and smallest when f_0 is 0. So, if you see that ΔF is largest when the contribution f_0 from the anomalous scattering atoms is $\pi / 2$ out of phase with that of the non anomalous atoms and ΔF is smallest when the f_0 contribution is in phase with that of the non anomalous atoms.

(Refer Slide Time: 33:22)

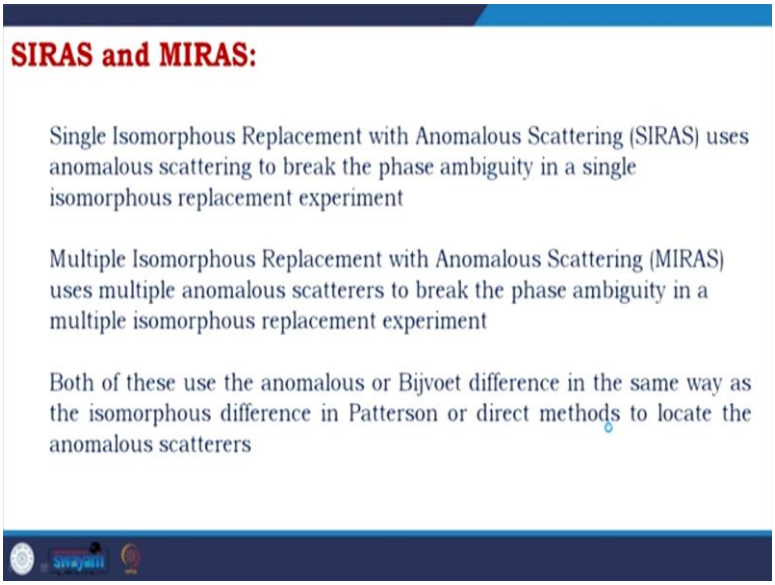


Variation in anomalous scattering at the K (peak) edge of selenium,

$f(\theta,\lambda) = f_0(\theta) + f'(\lambda) + if''(\lambda)$

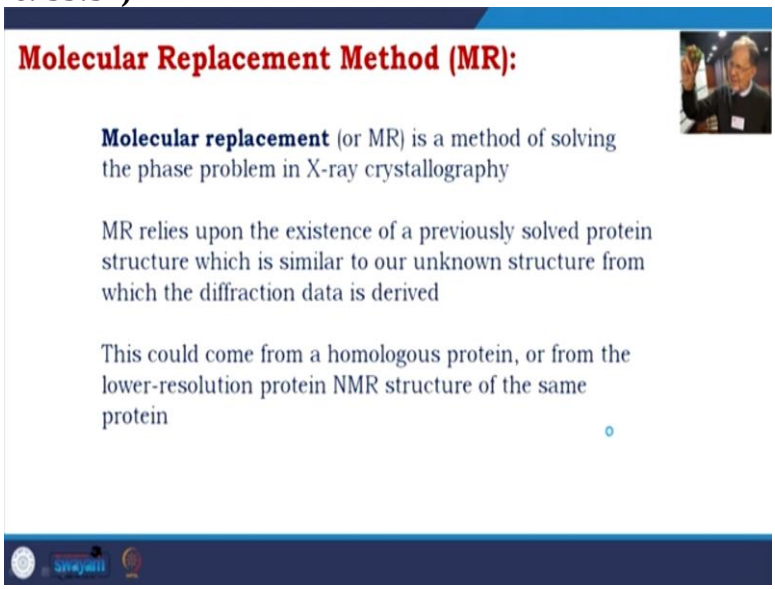
where a prime lambda is the dispersive term and double prime lambda is the absorption term. So, if you see here, the K edge is where the f' and f'' are greatest for selenium, the K edge is 0.9795 angstrom.

(Refer Slide Time: 34:05)



SIRAS is single isomorphous replacement with anomalous scattering uses anomalous scattering to break the phase ambiguity in a single isomorphous replacement experiment, you could consider it as a hybrid of both the isomorphous replacement along with the anomalous dispersion or anomalous scattering. Multiple isomorphous replacement with anomalous scattering is MIRAS, it uses multiple anomalous scatterers to break the phase ambiguity in a multiple isomorphous replacement experiment. Both of these use the anomalous or Bijvoet difference in the same way as the isomorphous difference in Patterson or direct methods to locate the anomalous scatterers.

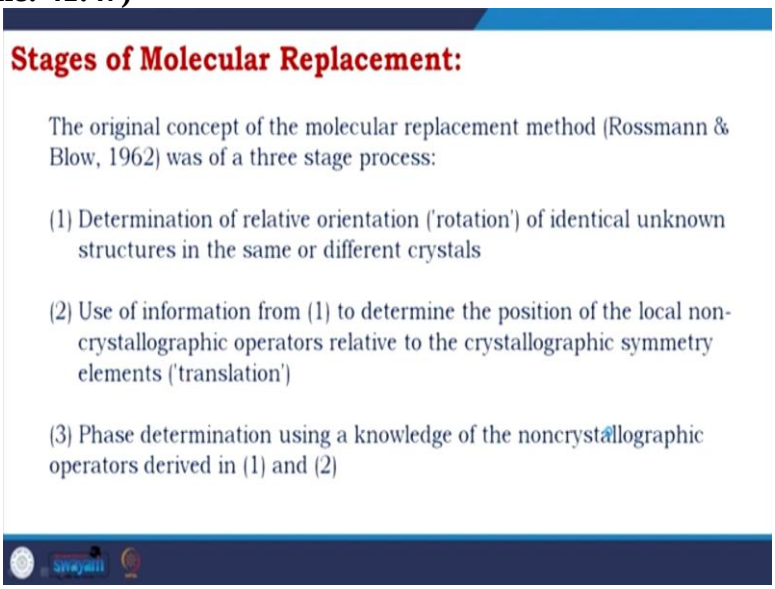
(Refer Slide Time: 35:31)



Molecular replacement (MR) is a method of solving the phase problem in X ray crystallography. This process was developed by Michael Rossmann. MR relies upon the existence of a previously solved protein structure which is similar to our unknown structure

from which a diffraction data is derived. This could come from a homologous protein which is having similar property to the protein you are working or from a lower resolution protein NMR structure of the same protein.

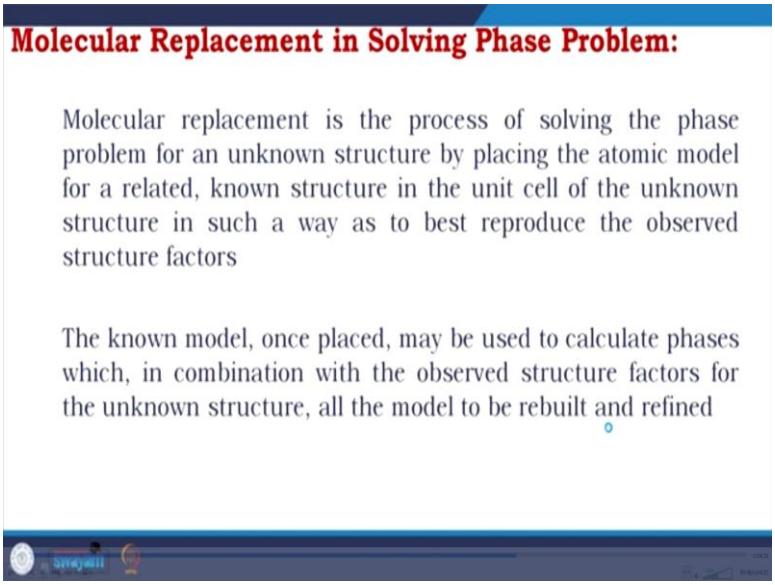
(Refer Slide Time: 41:47)



The original concept of the molecular replacement method was of a three stage process.

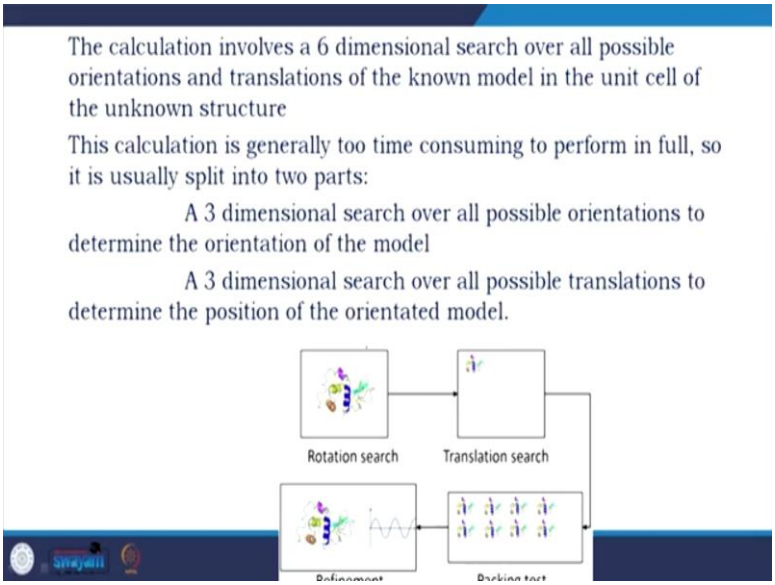
1. Determination of relative orientation (rotation) of identical unknown structure in the same or different crystal
2. use of information from one, to determine the position of the local non crystallographic operators relative to the crystallographic symmetry element (translation).
3. Phase determination using knowledge of the non crystallographic operators derived in 1 and 2. So, you do the rotation, you do the translation and you utilize them to get a phase.

(Refer Slide Time: 42:44)



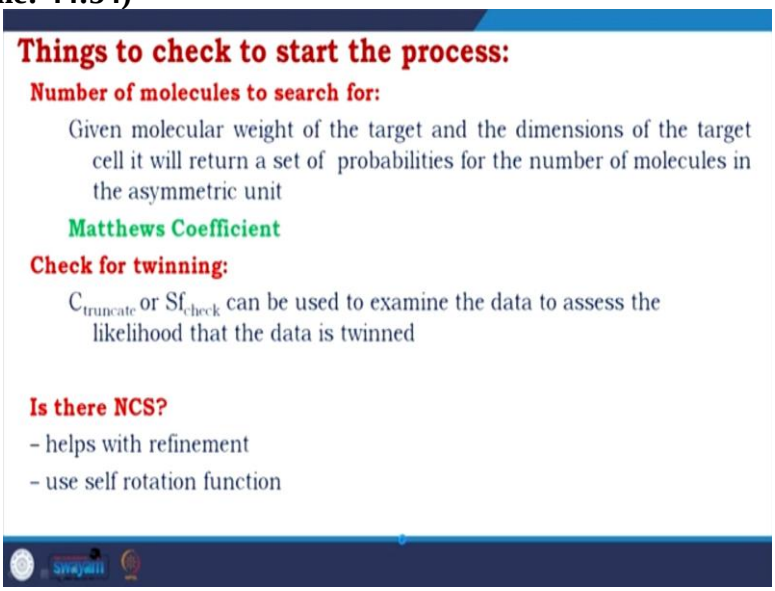
Molecular replacement is the process of solving the phase problem for an unknown structure by placing the atomic model for a related known structure in the unit cell of the unknown structure in such a way as to best reproduce the observed structure factors. The known model, once placed, may be used to calculate phases which in combination with the observed structure factors for the unknown structure, all the model to be rebuilt and refined. So, from the experiment, you get the structure factor, you get the amplitude. Somehow by doing the theoretical operation of rotation and translation, you feed that with the previous result structure and get the phase and solve the new structure.

(Refer Slide Time: 43:49)



The calculation involves a 6 dimensional search over all possible orientation and translation of the known model in the unit cell of the unknown structure. This calculation is generally too time consuming to perform in full, so usually it is divided into 2 parts, the first one, a 3 dimensional search over all possible orientation to determine the orientation of the model, and second, a 3 dimensional search over all possible translations to determine the position of the oriented model.

(Refer Slide Time: 44:54)

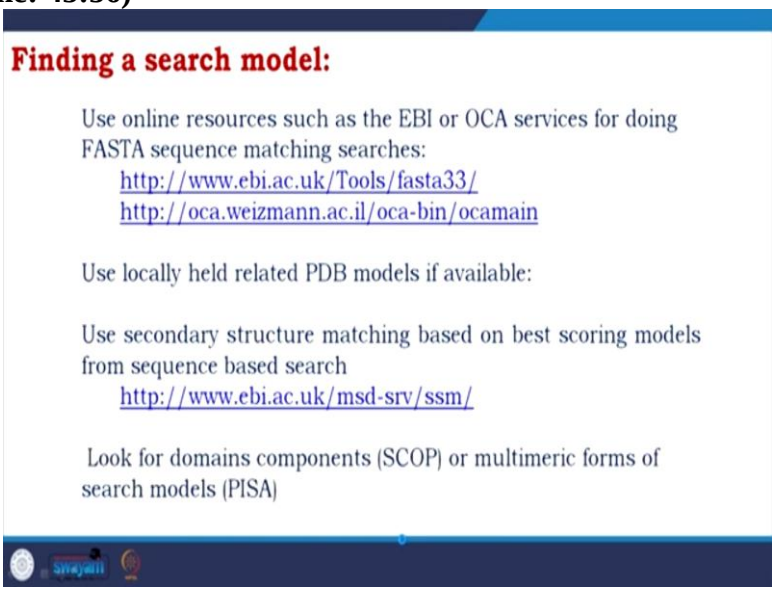


There are things to check to start the process of molecular replacement, one the number of molecules to search for. Given the molecular weight of the target protein and the dimension of the target cell, it will return a set of probabilities for the number of molecules in the asymmetric unit which is Matthews coefficient. We also have to check for twinning, there are

programs $C_{truncate}$ or Sf_{check} which could be used to examine the data to assess the likelihood that the data is twinned or not.

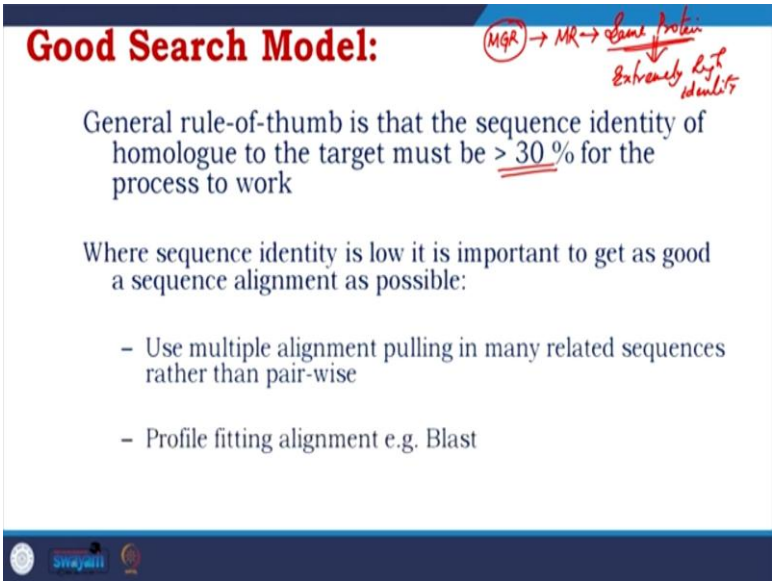
Is there non crystallographic symmetry? You also have to check, this information would be helpful with refinement and you could use self rotation function.

(Refer Slide Time: 45:56)



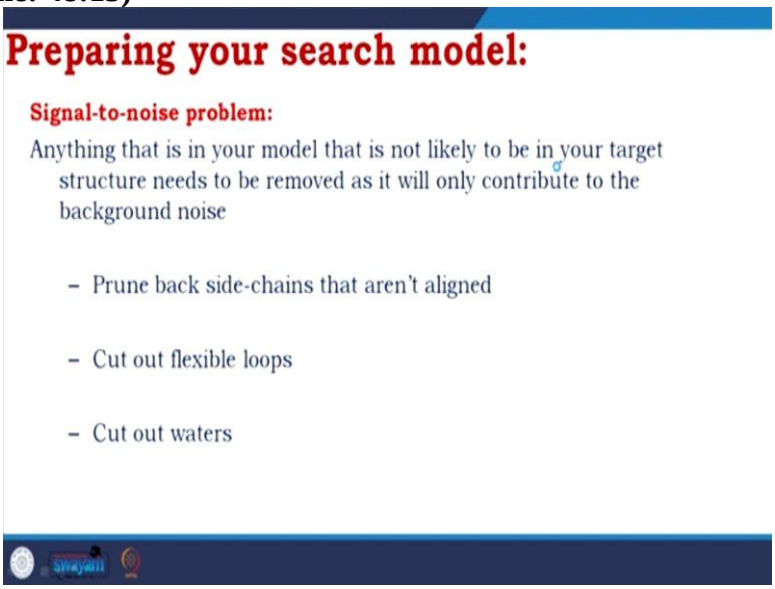
So, first we have to do a finding a search model. There are online resources such as the EBI or OCA services for doing, fasta sequence matching searches. You could also do locally held related PDB models if available. Suppose, you already solved one of your structure and you are making mutants, you could use the local PDB to do that. You could also perform secondary structure matching based on best scoring models from sequence based search. You could look for domain components from SCOP or multimeric forms of the search models from PISA.

(Refer Slide Time: 46:50)



General rule of thumb is that the sequence identity of homologue to the target must be greater than 30% for the process to work. Even when the sequence identity is lower than 30%, it is important to get a good sequence alignment as possible. You could use multiple sequences alignment pulling in many related sequences rather than pair wise, also you could go for global profile fitting alignment like BLAST.

(Refer Slide Time: 48:15)

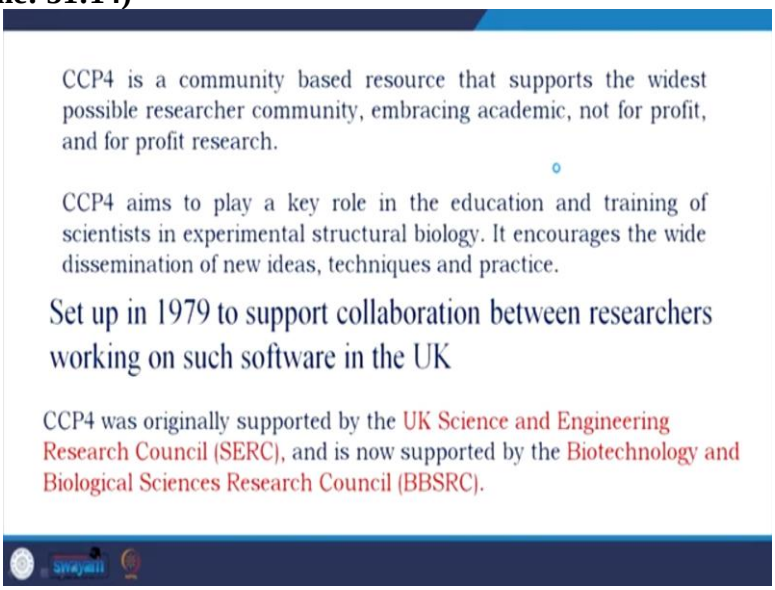


How to prepare your search model? Signal to noise problem, anything that is in your model which is not likely to be in your target structure needs to be removed as it will only contribute to the background noise. They have been very different in terms of binding ligands, solvent, different loops, very different flexibility in different part of the loops, side chains. You have to take care initially, cut out the flexible loops, crystallographic solvent or precipitant etc.

(Refer Slide Time: 50:16)

CCP4 is collaborative computational project 4, it is a developing software for macromolecules X ray crystallography. CCP4 exist to produce and support a world leading integrated suite of programs that allows researchers to determine macromolecules structures by X ray crystallography and other biophysical techniques. CCP4 aims to develop and support the development of cutting edge approaches to experimental determination and analysis of protein structure and integrate these approaches into the suite.

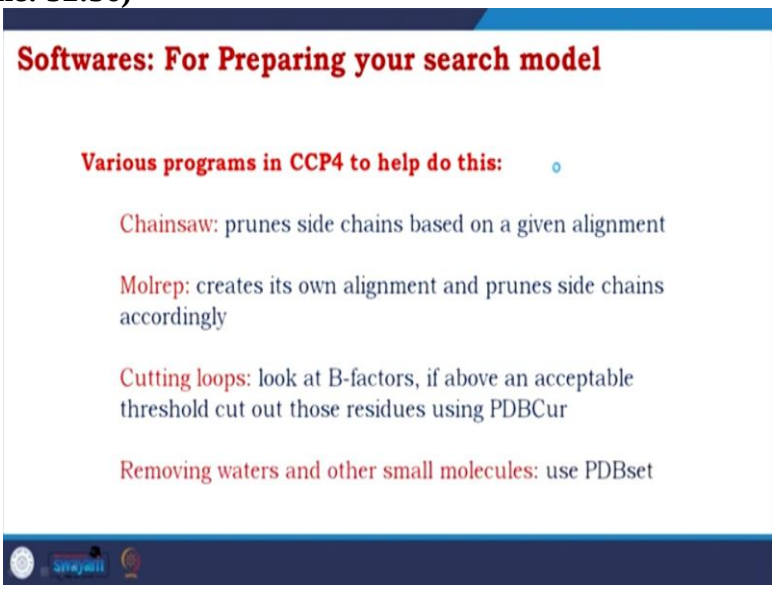
(Refer Slide Time: 51:14)



CCP4 is a community based resource that supports the widest possible researcher community, embracing academic, not for profit and for profit research. CCP4 aims to play a key role in the education and training of scientists in experimental structural biology. It encouraged the wide dissemination of new ideas, techniques and practice. Set up in 1979 to support collaboration between researchers working on such software's in United Kingdom,

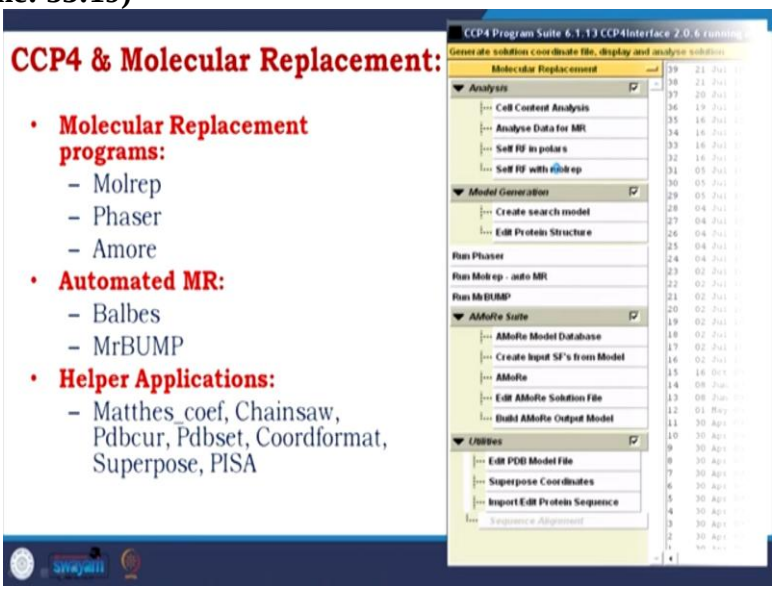
CCP4 was originally supported by the UK Science and Engineering Research Council which is a SERC and is now supported by the Biotechnology and Biological Sciences Research Council which is BBSRC.

(Refer Slide Time: 52:36)



We need software's for preparing the search model. There are various programs in CCP4 to help doing this. Few I am naming, if you go and play with the software you will know more. Chainsaw: prunes side chains based on given alignment. Molrep: creates its own alignment and prunes side chain accordingly. Cutting loops: look at B factors, if above an acceptable threshold cut out those residues using PDBCur. Removing water and other small molecules: use PDBset.

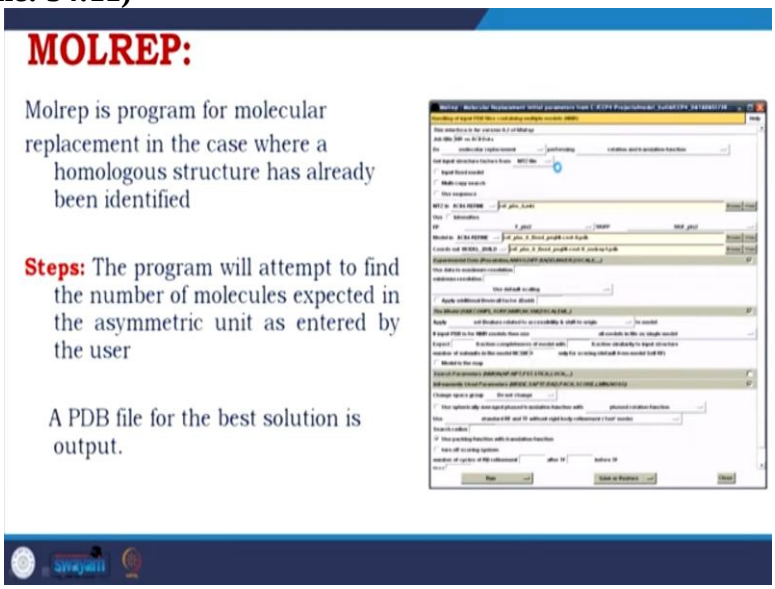
(Refer Slide Time: 53:19)



CCP4 and molecular replacement. There are molecular replacement programs Molrep, Phaser, Amore. Automated MR, Balbes, MrBUMP. Helper applications: Matthes coef,

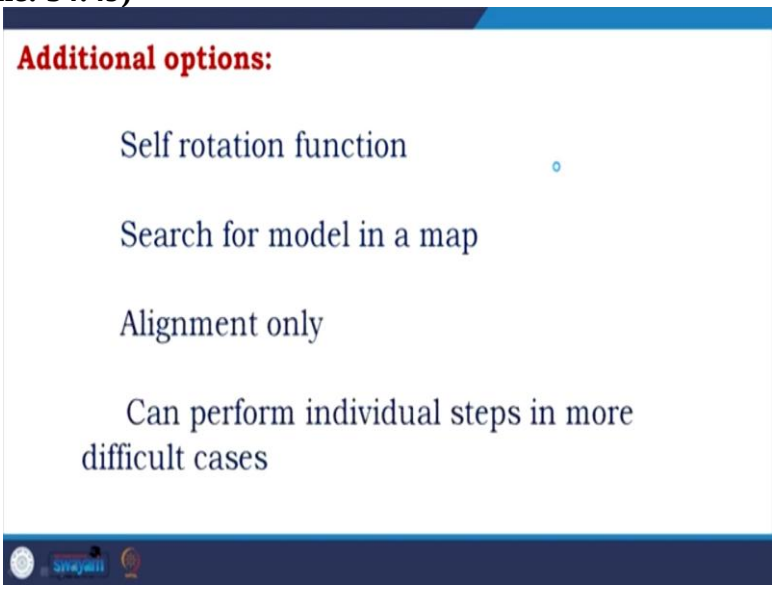
Chainsaw, Pdbcur, Pdbset, Coordformat, Superpose, PISA. So, this is the CCP4 program suite you could go and check whatever they have. They have categorized them in analysis, model generation, Amores suite, utility and in those different categories.

(Refer Slide Time: 54:11)



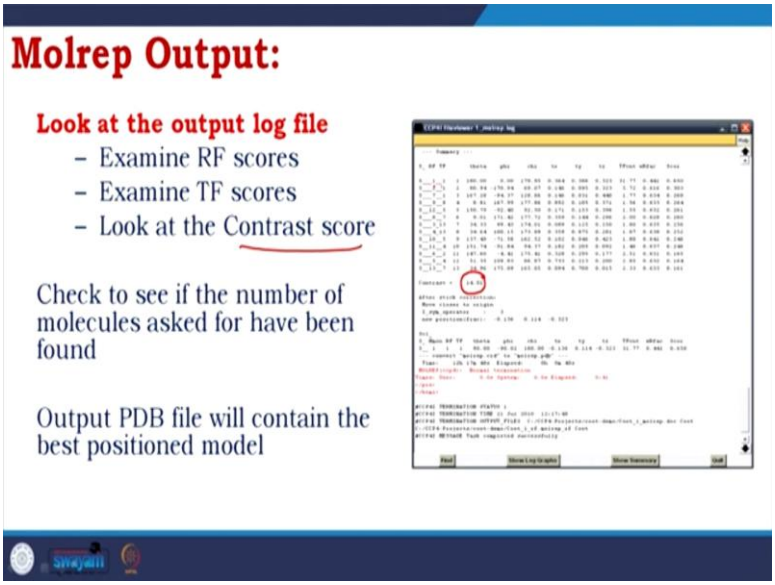
Molrep is program for molecular replacement in the case where a homologous structure has already been identified. The program will attempt to find the number of molecules expected in the asymmetric unit as entered by the user. A PDB file for the best solution is the output.

(Refer Slide Time: 54:49)



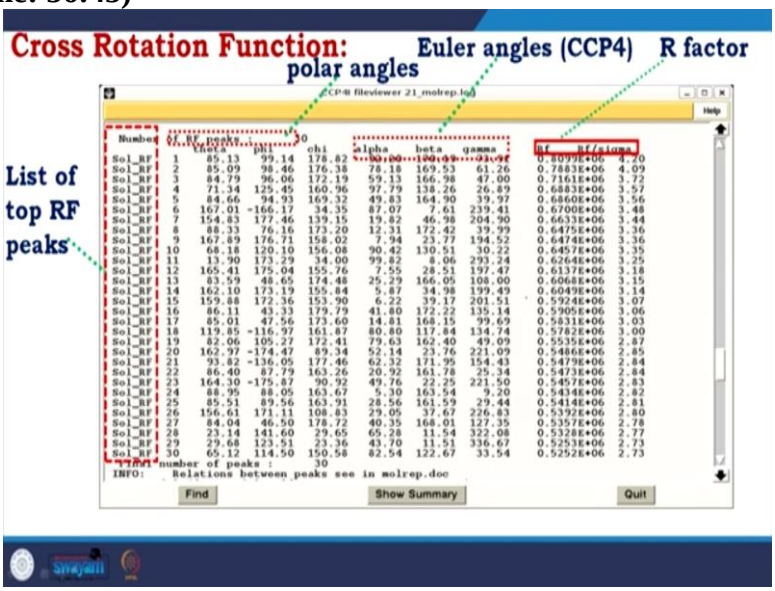
Additional options are there which self rotation functions are. Search for model in a map, alignment, also can perform individual steps in more difficult cases are the additional options in the suites.

(Refer Slide Time: 55:13)



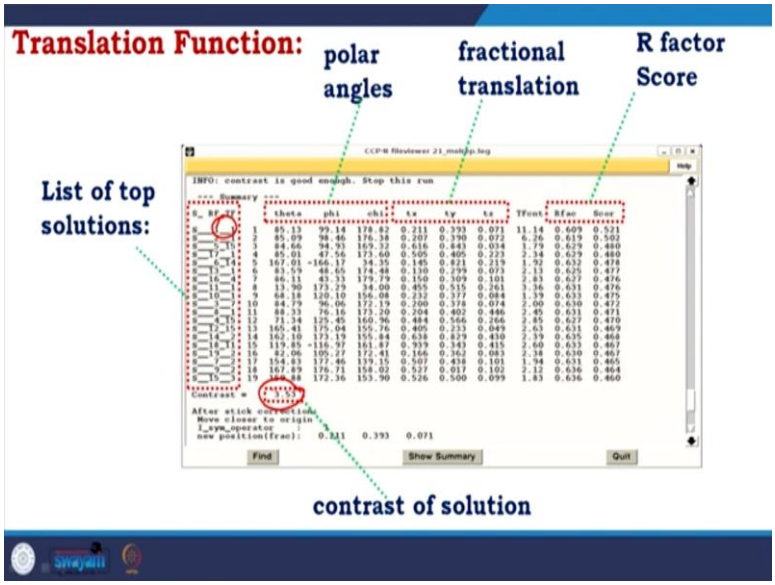
If you look at the Molrep output, you could look at the output log file, examine the RF (the rotational function score), examine the translational function score and also you look at the contrast score. Check to see, if the number of molecules asked for have been found. Output PDB file will contain the best positioned model.

(Refer Slide Time: 56:45)



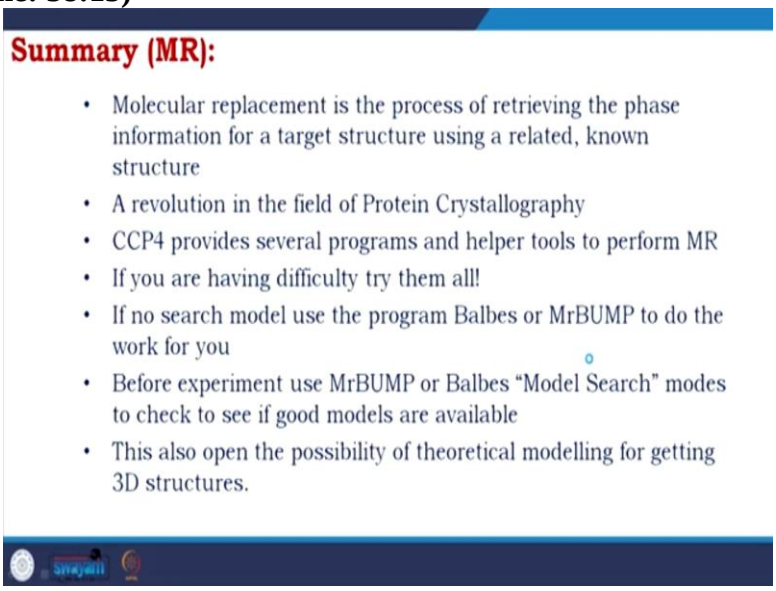
So, this is details about the cross rotation function, you will see the list of top rotational function peaks and you see how polar angles, Euler angles which is calculated specifically in CCP4 and you get the R factor. R factor, it will talk about the organizer the structure, we will talk in details in the next class.

(Refer Slide Time: 57:29)



In the translational function, you see the top solution, polar angles, fractional translations like translation, R factor, the score. More refinement you do, less the R factor would be, and that is one of the very good estimation of how your refinement is going.

(Refer Slide Time: 58:15)



In summary, molecular replacement is the process of retrieving the phase information for a target structure using a related known structure. It is a revolution in the field of protein crystallography. CCP4 provides several programs and helper tools to perform molecular replacement. If you are having difficulty, you should try them all. If no search model, use the program Balbes or MrBUMP to do the work for you. Before experiment, you could use MrBUMP or Balbes, model search modes to check to see if good models are available or not.

Thank you very much. We have kind of finishing up the X ray crystallography module. We have only one class left. And I will discuss about how to submit and refine the structure. And then after refinement, we will deposit it.