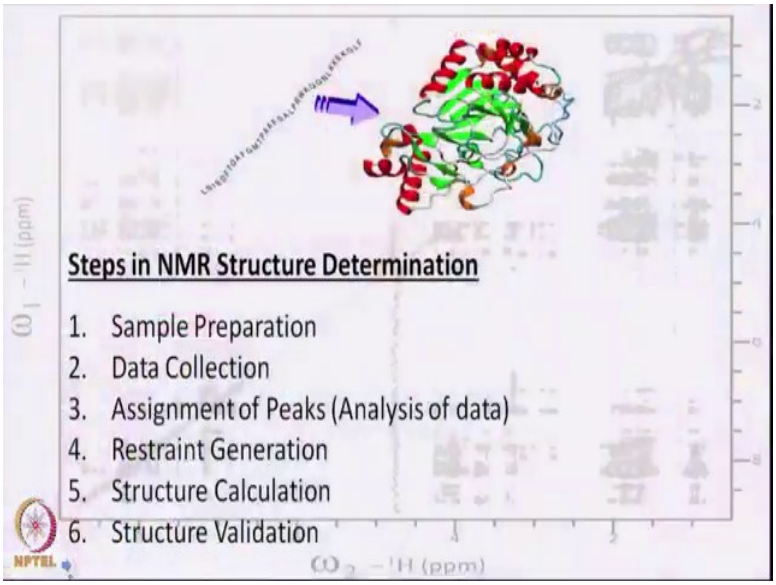
NMR Spectroscopy for Chemists and Biologists
Professor Doctor Ashutosh Kumar
Professor Ramkrishna Hosur
Department of Biosciences and Bioengineering
Indian Institute of Technology Bombay
Lecture 57
Structure Determination of peptides by NMR II

Hello, welcome to today's lecture. So we were discussing about structure determination of peptide by NMR spectroscopy.

(Refer Slide Time: 00:27)



So now, so essentially what we want, just to repeat whatever we did in the last class, we have a primary sequence, amino acid sequence, and then we want to translate that amino acid sequence how this polypeptide will fold in 3D structure and this for shorter peptide like 20 to 50 amino acids, how we can get the 3D structure using NMR spectroscopy as a tool. So that was what we were discussing.

So we discussed there are some primary steps in NMR structure determination. First is sample preparation. So for peptide we need a, we need to synthesize a peptide or bacterially express this peptide and purify it. The pure peptide has to be dissolved in a solvent which we want to solve this structure. And then 10% of D_2O is, or deuterated solvent is added to that for locking the spectrometer. So that is our sample preparation.

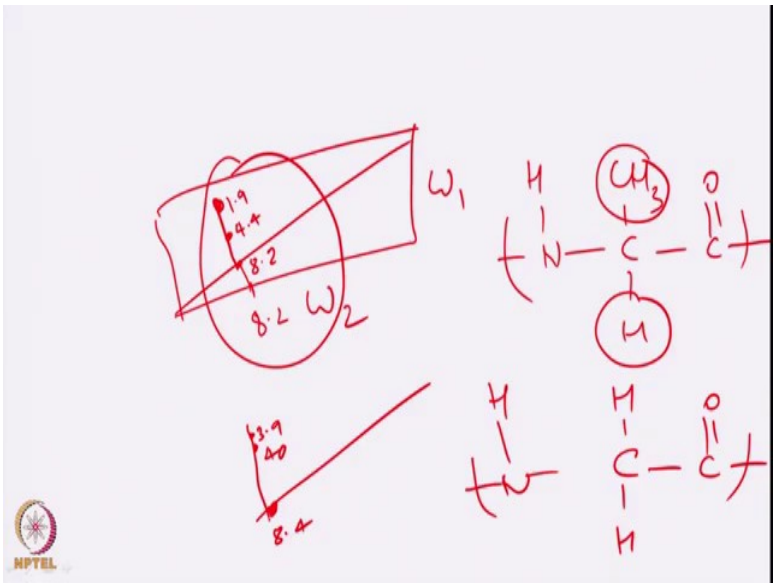
It has to be transperent. There should not be any, like aggregated particle so that we have to make sure that sample is absolutely pure and sample looks transparent. That says that it is pure. For purity we can use orthogonal techniques like mass spectroscopy, MALDI or even

SDS PAGE electrophoresis to see if the band is single, and in MALDI you are getting a single band and that tells about the purity of the sample.

The next step that we have to record a series of NMR data that we discussed last class so like TOCSY or Total Correlation Spectroscopy and NOESY, Nuclear Overhauser Effect Spectroscopy, then couple of like ^{15}N HSQC, ^{13}Carbon HSQC or Double Quantum Filter COSY. So all these data set we need to record it and that takes a day or may be 2-3 days and after that we were discussing actually the assignments of the, how we are going to interpret this data. So we ended in the previous class at the analysis of the data.

So now we will discuss that and today we are going to discuss more about how you use this data to generate the restraints and those will be used for calculating the 3D structure and once the structure is calculated then we need to validate that structure. Does it satisfy the Ramachandran plot statistics and all those and then it has to be minimum energy structure. Once it is validated and energy is; all the angular constraints and the distance constraints are satisfied then this structure is good for submitting to protein data bank. So that is what we will be discussing today.

(Refer Slide Time: 03:16)

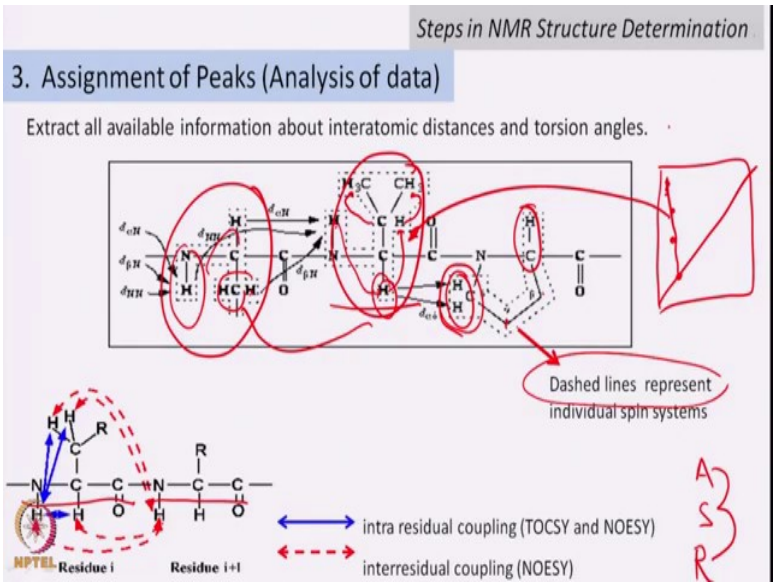


So we will be discussing about, let us say simple thing that I was showing you, assignment of few amino acids. Like we just showed it for alanine, so let us see this is our alanine in a polypeptide chain. So this is alanine and let us write it down for glycine, that is what we had written here, glycine, okay, so now we were looking how TOCSY pattern will come. So here is our, say spectrum, ω_1, ω_2.

So around 8.2 ppm our NH peak will come. So here is my NH peak. Then we have a peak coming H_α. So we will get a peak from H_α and then we will get a peak from H methyl. So if this pattern and this will be say around 1.9, this will be around 4.4 or something like that, this will be 8.2. So if we get this kind of pattern in a TOCSY spectrum we now, we understand that this is alanine.

Similarly for, say glycine we can have around 8.4 or so and then we have 2 closer peaks around 3.9 and 4. So this is pattern for glycine. So these amino acids have a typical pattern in the TOCSY spectrum and that we can use to identify in the, in the TOCSY spectrum and that will be our start point or something. Then we need to combine this with NOESY spectrum to walk around the backbone.

(Refer Slide Time: 05:05)



So here, I am giving you for analysis of data. So now we take a valley. Here, for valley what we have? We have NH. NH will be correlated with say, H_α. H_α will be correlated with H_β. H_β will be correlated with methyls. So if we, in a spectrum, in a TOCSY spectrum we get pattern for this. Suppose I write it down. So suppose here is our NH. And we get a peak for H_α which is 4 point, say 5, then H_β1.8 and then H_γ1 and H_γ2, if very close. So if you get peaks something like this, this we know that this is coming from valine

Similarly we can get a pattern, say for any of the amino acid like proline, so proline we do not have NH, right, so here that is missing. So we have here 2 peaks and can be here and all those, so here also. If we get a pattern like this we know that this is coming from proline.

Similarly we can take it from some of these here, so NH, that will correlate with H_α, H_β and so on and so forth.

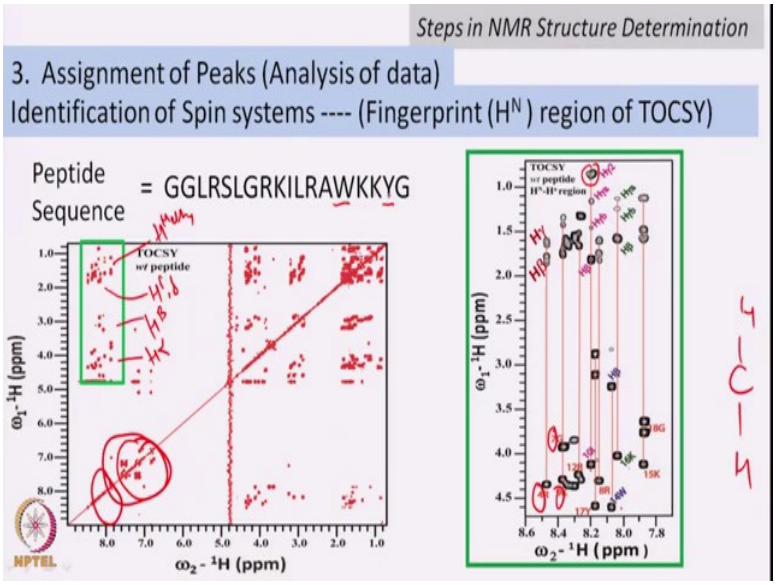
So these will be coming for a simple spin system, each of these spin system will give a typical pattern, that is with TOCSY, Now we want to see which is near what? So then we need a, like distance based spectroscopy and that is NOESY and if we do NOESY, we know that this CH_2 can show a correlation peak with this NH.

Now if this correlation in the NOESY spectrum from NH to CH_2 we will see then we know that this valley is near to this residue. Similarly if we see some H_α showing correlation peak with proline, this protons then we know that this is coming, these two are near in space and that is how they have a, they have the correlation peak in the NOESY spectrum.

So that is how we try to identify from the TOCSY and NOESY, how, which residue are like, how they are connected. So here just for schematic we have shown two residue, *i* and *i+1*, so this is, this one residue is *i* up to here and this is *i+1*. So in TOCSY you will get all the blue correlation which is like intra-residueNH2CH alpha and then 2 methyls, right. And in NOESY you can get same, like in the same amino acid as well the neighboring amino acid.

So inter-residue as well intra-residue correlation we can get it, okay. So here intra-residue means this CH_α will be correlated with NH of next residue. This NH can also be correlated with the NH of the next amino acid and that that helps us in identification of the proximity residue. So in TOCSY we know that this is alanine, this is serine, this is arginine also but if we combine NOESY with this, we can now, so this alanine, serine and arginine are near in this space and that is how we can identify.

(Refer Slide Time: 08:29)



So I will walk you through some of those assignment protocol. So let us see. I am taking a sequence of 18 amino acid which is GGLRSLGRKILRAWKKYG. So this is single letter code of different amino acids and this is say, one peptide which is 18-mer. Now we have a beautiful TOCSY spectrum. So at the moment let us for simplicity I just concentrate on this region which is NH; H like H aliphatic so these are all H_α, H beta and here are H methyls.

So say, let me write it down, H_α, here will be H_β and H_γ and δ, right ϒ, δwhatever and here are H methyls. So only I am focusing on this region. So let us take, here are my first pattern. So for arginine, here is H_α, H_β, β 2, ϒ. Now for lysine we have 1 here and then series of here. And then you see intensity coming from glycine.

Now here glycine both the H, CH, H these H_2 glycines are showing peak at the same position because their chemical shift is not too different and that is how we have only one peak for that. Similarly you can find again for arginine, here is for isoleucine H_α, H_β, $H_\gamma 1$, ϒ2 and similar all the way of methyl. Now for say tryptophan here, tryptophan we have H_α, H_β. Lysine you have all those peak, H_α, H_β, H_γ, H_δ, so, and again all those for these.

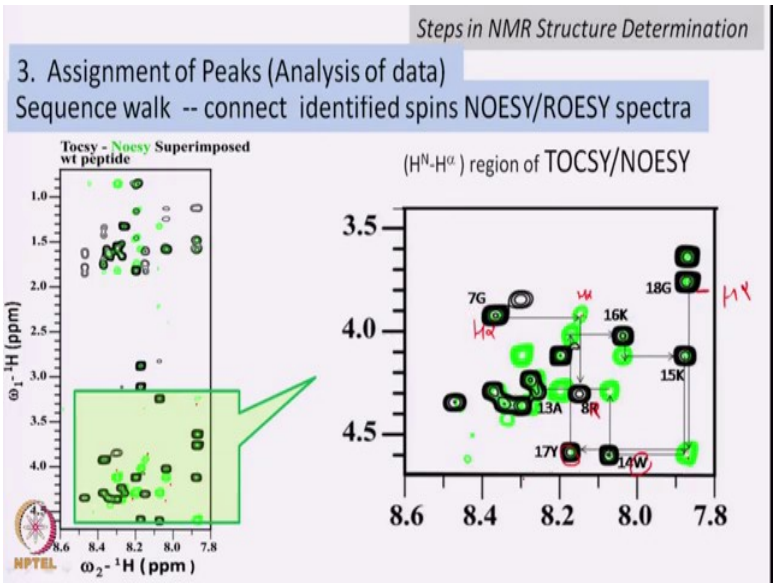
So different amino acid type now looking at this TOCSY pattern, zoom it at TOCSY pattern, we can find it out which residue it is. But like this 7, 4, 9 does not only come from TOCSY. For that I need I need to something called NOESY spectrum. We combine it and in next slide I am going to show.

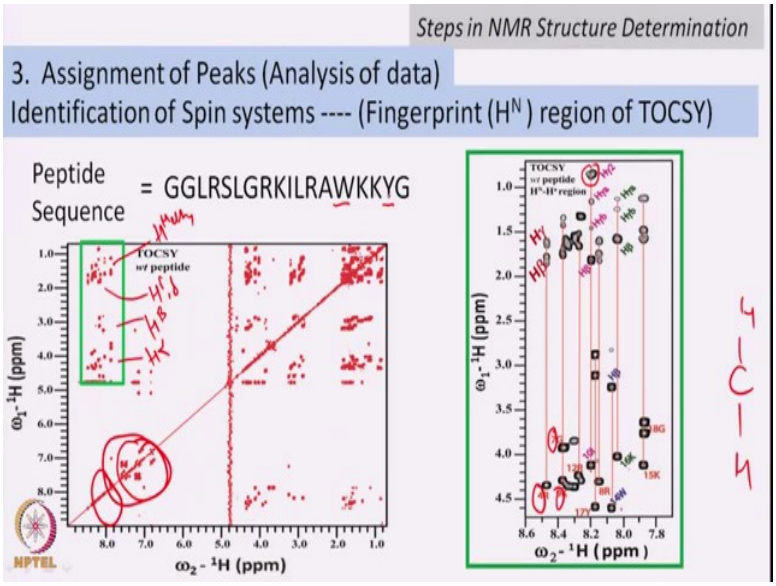
But other than that if you zoom this region, this region is correlation of NH and NH. Like in each amino acid at least one NH is there and this, this population here shows that NH and NH are connected. This gives very important structural information. Now as we have studied NOESY gives you distance dependent correlation.

So that means this NH--NH correlation if it is shown in NOESY spectrum, this is TOCSY spectrum but I will come to NOESY spectrum, it is shown in NOESY spectrum then that gives information about the distance. Here is the TOCSY spectrum. If we this, we have a NH, NH correlation that clearly says that the amino acid which has NH as a backbone as well as side chain can also give you here. But this is also like, if you look at 7.2 to 6.7 this are from aromatic amino acid. Now aromatic amino acids are very important.

So if you look at here, YW these are aromatic amino acids and then they can give correlation of the aromatic side chain correlation and that is shown here. So NH---NH correlation will be shown here in the NOESY spectrum that I will show you. This is aromatic protons correlation that comes around this and this is coming from tryptophan and tyrosine. So let us combine this TOCSY information with NOESY spectrum.

(Refer Slide Time: 12:07)





So here what I show for the same peptide, 18 amino acid peptide, TOCSY is shown in black and NOESY is shown in green. So if you see NOESY as we discussed it will give you intra-residue correlation as well as inter-residue correlation. Now intra-residue will be same like TOCSY but inter-residue will be different. So if you look at here, many of the green peaks are also similar to the black peak. So this is coming from intra-residue.

However these all peaks are coming from inter-residue correlation, not intra-residue. The green-black overlap is intra-residue. The green alone is inter-residue correlation. So now this green, if you identify this green because this is distance dependent or a dipolar coupling based correlation so we know which residue is near to what and that is what helps us in identify. And again this is NH, H_α section. So if we zoom this section, same thing we have here.

So let us start from anywhere. Now say here, let us start from glycine 18. As we said glycine H_α will be, will be very intense and this coming at around 3.7 or 8 ppm, so we know that this is glycine. Now glycine H_α is showing a correlation peak with another amino acid here. Now, so you see this is H_α. So H alpha, H_α correlation from glycine 18 to something is nearby. Now what could be nearby?

It can be either 17 or 19. Now if you take this peak and go in this line, now here the peak and this peak has another correlation and that we have identified. So this is Y. So this has a correlation which can be traced back to the Y kind of residue that is tyrosine residue. So now we know that Y and G are near in space. So we will go back in the sequence and then try to

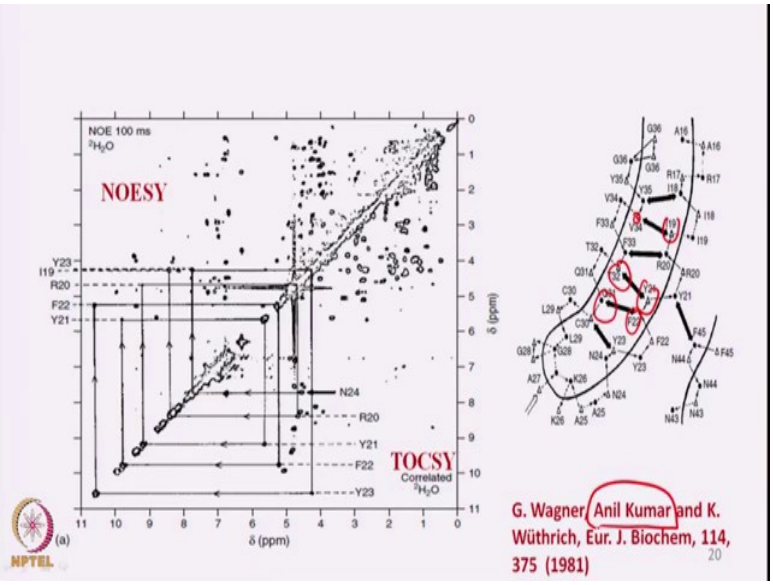
look at where are our Y and G? So here, if you, here Y and G, so now by using this we can identify this is 18 and Y is 17.

Similarly let us go to some other, here. Let us see, here what we have is 15 K. At the moment I do not know whether this is 15 or not, but I know this is coming from a lysine residue. Lysine also shows a correlation peak at this position, both of these are overlapped. Now lysine will show a correlation. Now we go back here and we find there is a residue which can be identified as tryptophan. So now tryptophan and lysine seems to be together and then we can look at here. So tryptophan and lysine are together, that means this is 15 lysine, this is 14 tryptophan.

Similarly we can do exercise for few of more amino acid. Like here if you look at glycine so here we had glycine, glycine's H alpha showing a correlation peak of H alpha and 8 H showing a correlation peak 8 R is showing a correlation peak which is this. And this shows correlation peak with this glycine. So now we know that arginine is showing correlation peak with a glycine type and arginine and glycine are nearby here. So that is 8 and 7.

So similarly we need to do exercise for days and we can walk across the backbone to identify each of these peaks using TOCSY and NOESY spectrum. This takes time. There are, as you see there are several overlaps. There will be several ambiguity. So by doing like we can then take, make use of this region which is NH H beta region and by using whole spectrum we can identify peak coming from the TOCSY and NOESY how they are correlated by bond and by space respectively and that is how we can identify amino acid number like 7 G, 8 R, 13 A or so and so forth.

(Refer Slide Time: 16:32)

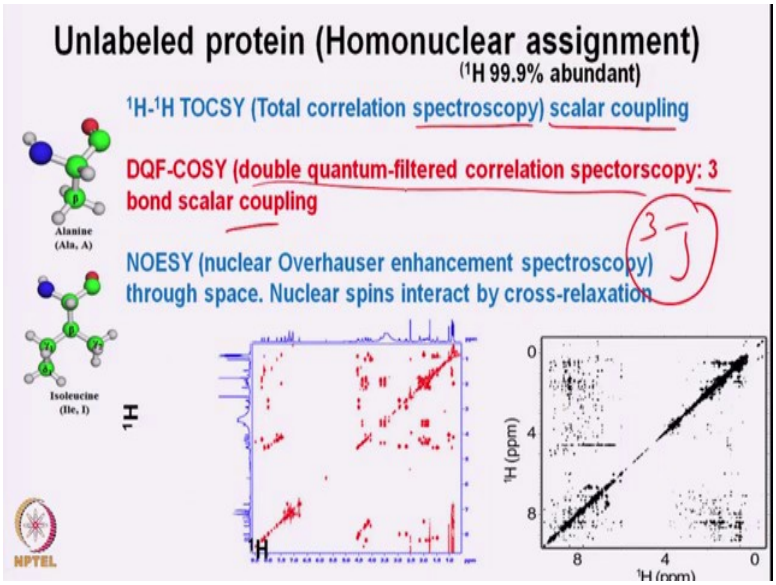


So this is also like, you can find it out and that helps us in building a model. So now I identify that if you look at the NH, H_α correlation is happening in backward, right, backward means 8 is showing correlation with 7, 13 is showing correlation with 12 and 18 is showing correlation with, with 17. And that is how we can look at, like these kind of correlation we are seeing here. So this helps us in building a model how the NOESY peaks will be assigned.

Similarly if we see a long range NOE constraints like for example from here to here, then we know that these two amino acids are closer in space and therefore by finding several kind of these correlation one can build a model for the polypeptide chain that my polypeptide chain will folded in such fashion, okay and that actually, essentially it can be identified using TOCSY and NOESY and this was very established technique like 30 years ago, almost 30 years ago.

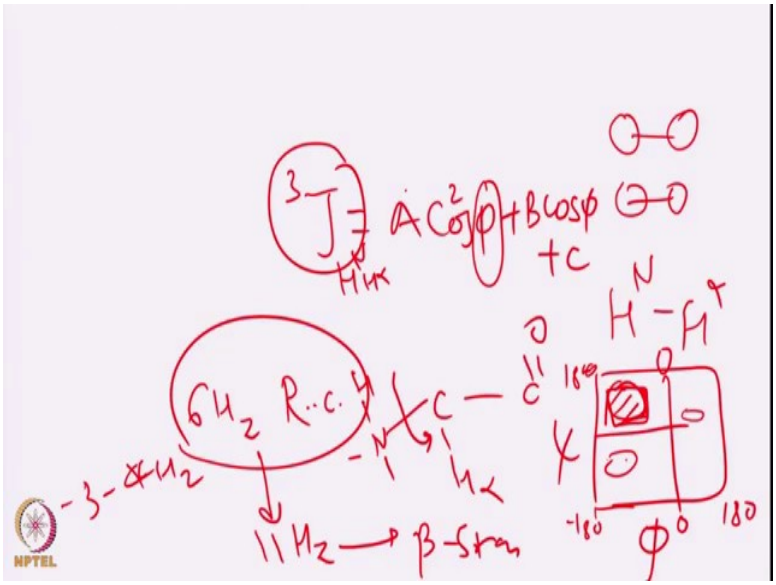
Professor Anil Kumar is a pioneer who has developed this NOESY, NOESY techniques and now is a core of the structural biology. He showed, actually for the first time how you can use NOESY for the structure calculation for polypeptide chain. So now this building model helps us in identifying and fixing up the assignments, great.

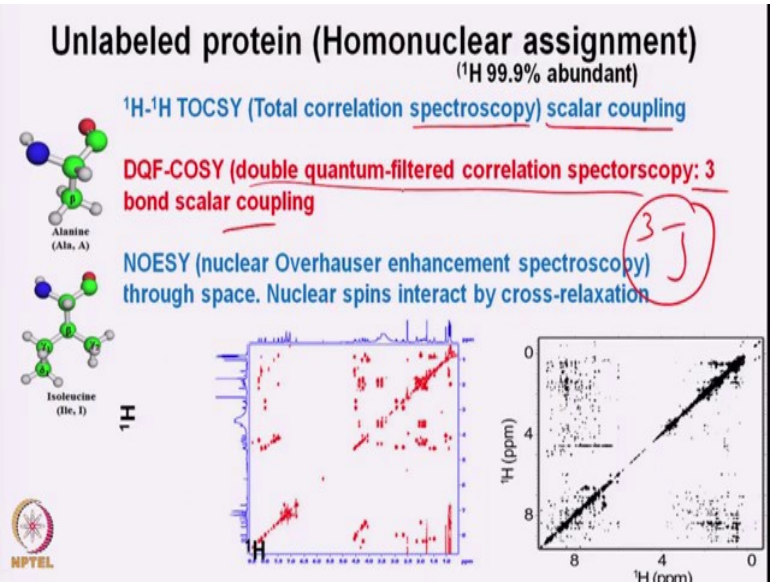
(Refer Slide Time: 18:02)



So that is what we discussed. The TOCSY total correlation spectroscopy which correlates, which correlates by scalar coupling can be used for intra-residual correlation and NOESY Nuclear Overhauser effect Spectroscopy can be used for, like intra-residue correlation. And then if we combine with something with Double Quantum Filter COSY, this gives you this 3-bond correlation, 3-coupling, ^{3}J, and if you remember we have like, we told that this ^{3}J is correlated with the phi torsion angle using the Karplus equation. So, just in a moment I will just give you some hints for that. So DQF COSY, you can essentially measure coupling constant, right?

(Refer Slide Time: 18:52)





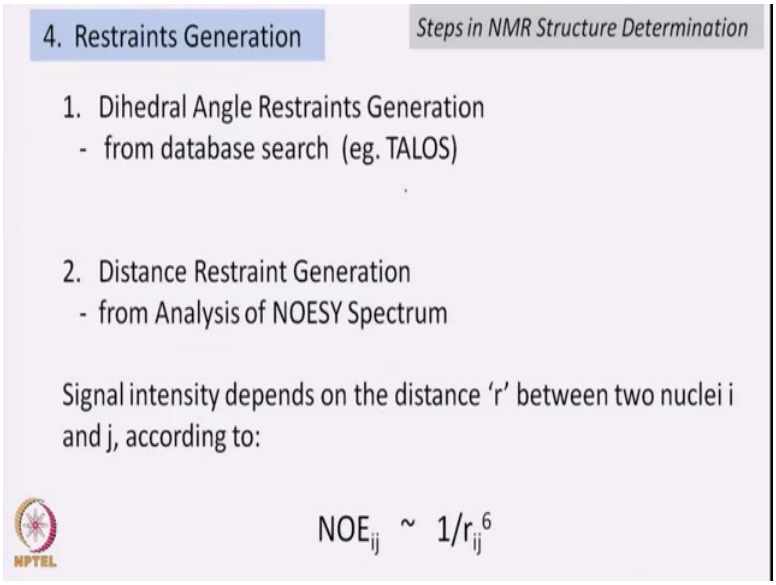
By looking at this splitting pattern you can measure something called HN--$H_\alpha$coupling constants so if you remember where are the H NH_α in polypeptide chain? This is essentially HN-- H_α and this gives you torsion angle. This, so if you remember your Ramachandran plot phi and psi, so this gives you phi value. So ^3J, is Karplus equation

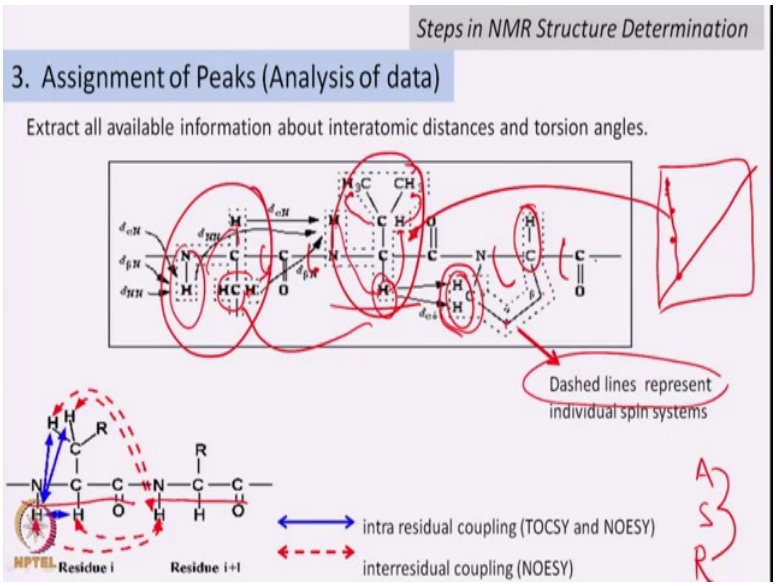
$$3J_{HNH\alpha} = A\cos^2\varphi + B\cos\varphi + C$$

So if you know ^3J, which comes from HN--H alpha, experimentally determined HNH alpha, you can calculate the phi torsion angle. So here is the phi torsion angle. So you know this is β strand region and this is φ_α region and this is L region, three permanent region so if, so just by knowing this J value one can know where my polypeptide, what kind of conformation my polypeptide chain has and say, for a random coil say typically it is a 6 Hertz or so, so 6 Hertz is for random coil chemical shift which does not have any persistent structure, if this ^3J, moves towards 10 to 11 Hertz they we are in beta strand region. And if it goes like 3 to 4 Hertz then we are in alpha helical region.

So just by looking at the J calculated for this Double Quantum Filter COSY we know where my polypeptide chain is. So suppose we are getting this ^3J, around say 9 to 12 Hertz then my polypeptide chain is in this region. That means if you remember the value, here is 0, this is 180, -180, +180 okay, so one can calculate what is the φ and ψ just by looking at, what is phi just by looking at the value J. So experimentally determined from DQF-COSY, Double Quantum Filter COSY you can experimentally determine and you can guess the conformation of polypeptide chain already.

(Refer Slide Time: 21:05)

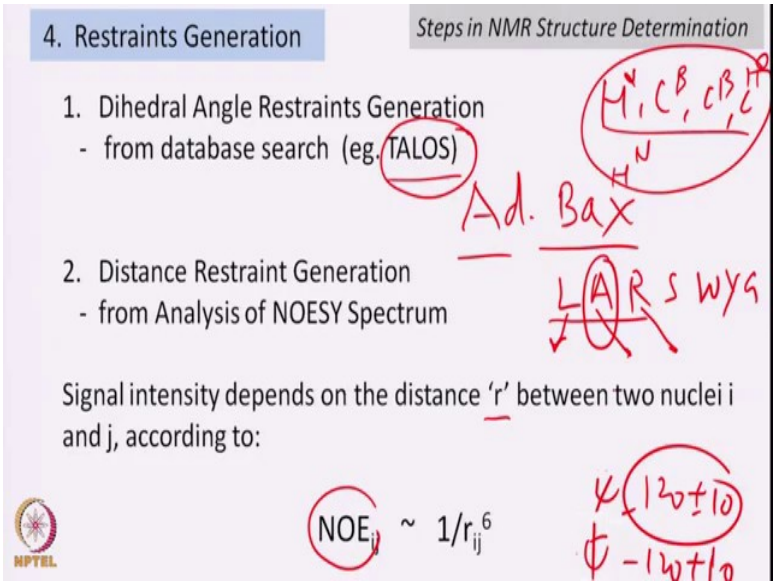


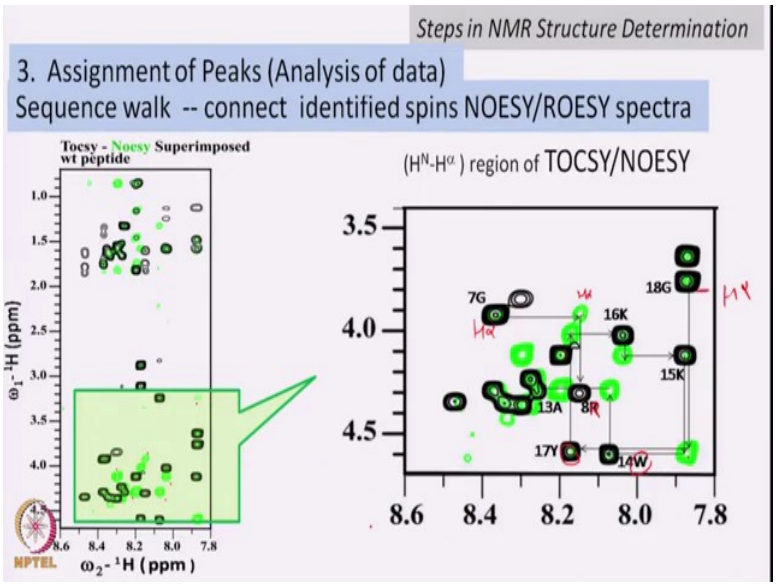


Now so, now we have got the assignment of peaks, we have got through-space correlation, we have got some idea about dihedral restraints coming from the DQF-COSY. So that all we will use in generating the restraints. So these restraints now can be further used for calculating a structure. So two things are important. One is torsion angle, how our polypeptide chain, what kind of conformation I had and then another is distance, distance. How these protons in space are correlated. If you go back here the correlation that we have shown, how these will be correlated?

What will be the distance from here to here, here to here, here to here, here to here? So if you get this distance and we get the torsion angles, these φ and ψ, φ and ψ then one can find it out the conformation of the polypeptide chain. So these two things we need.

(Refer Slide Time: 22:11)





Now over the years people have developed algorithm which is called TALOS, Torsion Angle Likelihood Obtained from the shift, chemical shift and that can be used for calculating the phi psi dihedral angle. So this was developed by Professor from NIH, Ad Bax, his group developed this algorithm called TALOS or TALOS plus which uses chemical shift that you have assigned from TOCSY and NOESY to predict the φ and ψ torsion angle for each of the amino acid.

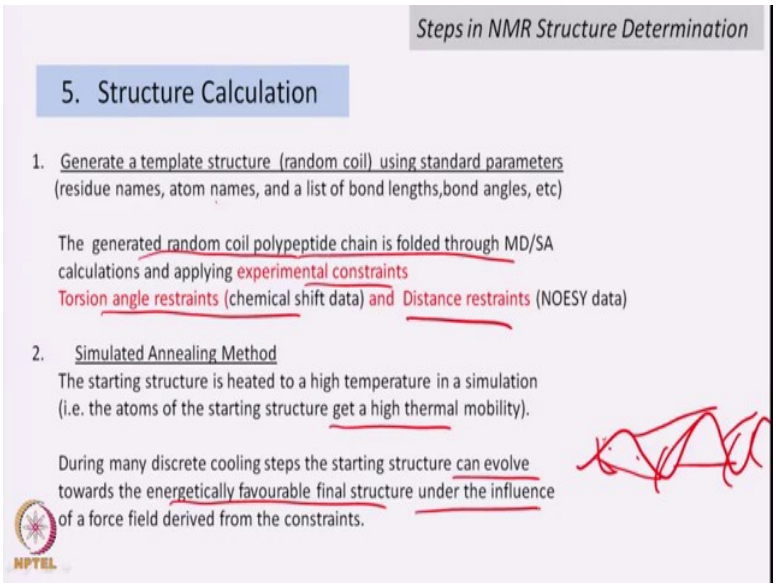
So what it does basically, TALOS, actually TALOS takes the chemical shift for 3 amino acids in a polypeptide chain. Suppose we have LARSWYG a something like that. So suppose I want to calculate the phi torsion angle of this alanine then TALOS needs 15 chemical shift

as input, 5 from here, 5 from here, 5 from here. What all chemical shift it needs? So it needs H_α, C_α, C_β, CO and HN.

So these are 5 chemical shifts it requires and then what it does, it matches with the database. Like already so many protein structures have been determined. So now so it matches this chemical shift in the database and also the peptide sequence in the database and then from the search it gives you information of these phi, psi that were there for all these structure determined. And then it predicts approximately with a range, say 120 ±10, this should be the say, for ψ this is 120, for φ is 120 ±10, something like this it predicts and gives you value.

So that you can use for a calculation of the torsion angle. Now as we have seen that NOE intensity from each of the peak that has appeared depends upon the distance *r*. So if we determine the intensity of these peaks that we had earlier here like, now we know that these peaks come from the correlation of which two protons, from two different amino acids we can find it out the distance of those amino acids. So intensity of NOE between *i* and *j*th peak depends upon $\frac{1}{r^6}$ and that gives you actually distance correlation.

(Refer Slide Time: 24:41)



So now you can use these dihedral constraints obtained from either DQF COSY or obtained from TALOS. Then you can use the distance constraints coming from the NOESY peak intensity and then one can generate a template. Template is essentially a random chemical shift.

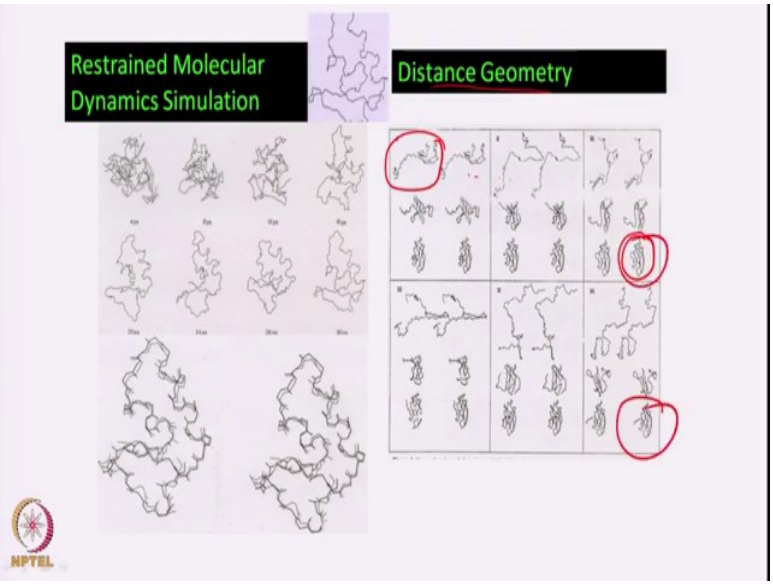
So residue name, atom number, then list of band bond length, and bond angle and then one can generate a random polypeptide chain. This does not require anything and then one can fold through molecular dynamics and then what we are doing, so we generated a polypeptide sequence like this, so like using the chemical shift and all those. Now we have angular constraint. What should be angle here? All these angles and we have distance from here to here, here to here, here to here all these distances we have.

So that is what you need. You need a template, you need an angle that will be required for folding and you need a distance. If you use these experimental constraints coming from the torsion angle restraints, coming from chemical shift data obtained by TALOS or DQF COSY and distance restraints from NOESY data then we are all set for structure calculation.

So what we do something called simulated annealing. So we start structure and then this structure is heated to high temperature in a simulation. And then atoms of the starting structure can get a high thermal mobility so everything that high thermal mobility is quite flexible. And then you do, you cool it in a stepwise and then these constraints will be, restraints will be used given by experimental constraints, so first you heat it then in an unbiased way slowly you cool it.

Now structure calculation algorithm will use these restraints and then they will minimize the energy and then energetically favorable final structure will appear and that used by this whatever force field we are using. So energy minimized structure will come by this simulated annealing process.

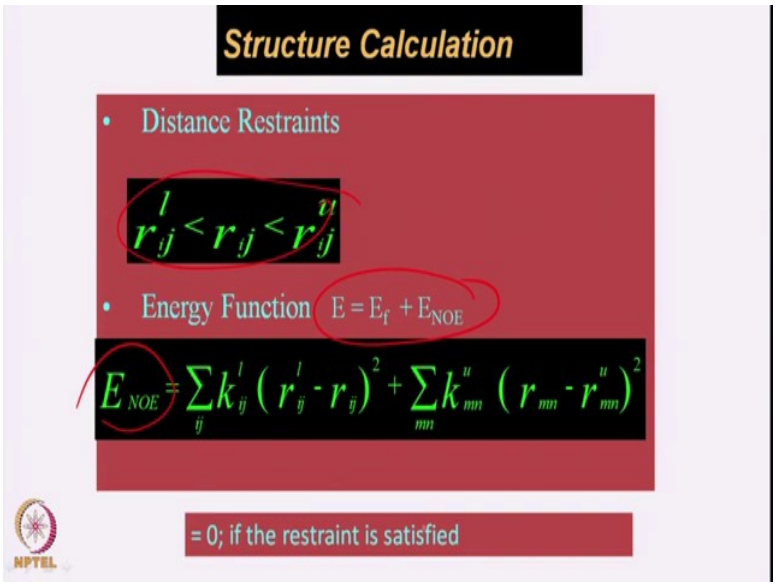
(Refer Slide Time: 26:50)



So here one can see, so you do restricted molecular dynamics simulation. You get a template and then you start, so like distance geometry, start from extended polypeptide chain, heat it at top and then slowly let it cool down so structure will be involving in a step. Here you can see at the end, finally we will obtain a folded structure.

So similar thing is shown here, starts from many conformation at high temperature slowly by using these restraints obtained from experiments and then making correct kinds of contexts here, the tertiary and secondary constraints. Finally a structure evolves and gives us a folded structure. So that is how you determine the structure of a polypeptide chain.

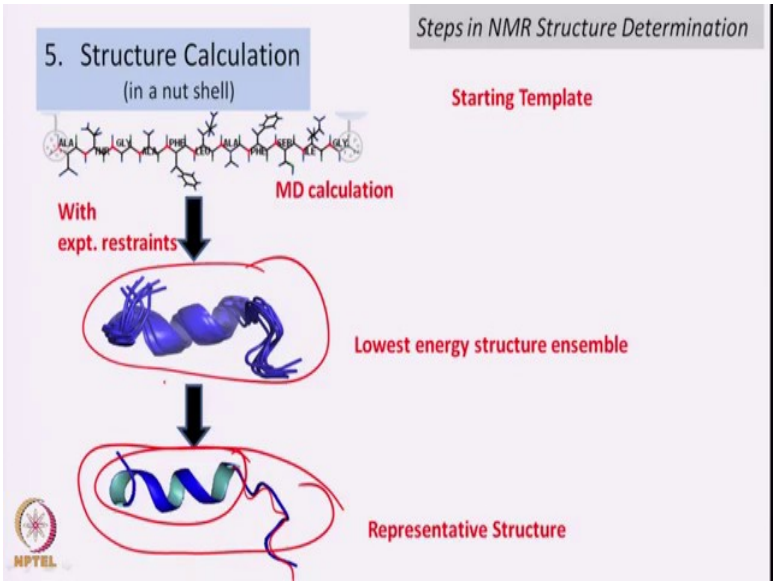
(Refer Slide Time: 27:39)



So essentially it solves some of those equations which minimizes the distances and it minimizes the energy. So by using this NOE energy, finally by minimizing the energy does the, it gets the minimum energy structure. So one thing to notice here NMR never gives you unique structure. It gives you ensemble of a structure, those which have minimum energy.

So the, if you look at any of the protocol that you use either CYANA or Xplor-NIH that gives you either 10 on somebody's structure or 20 on somebody's structure. These are minimum energy structure obtained by distance geometry simulated annealing protocol.

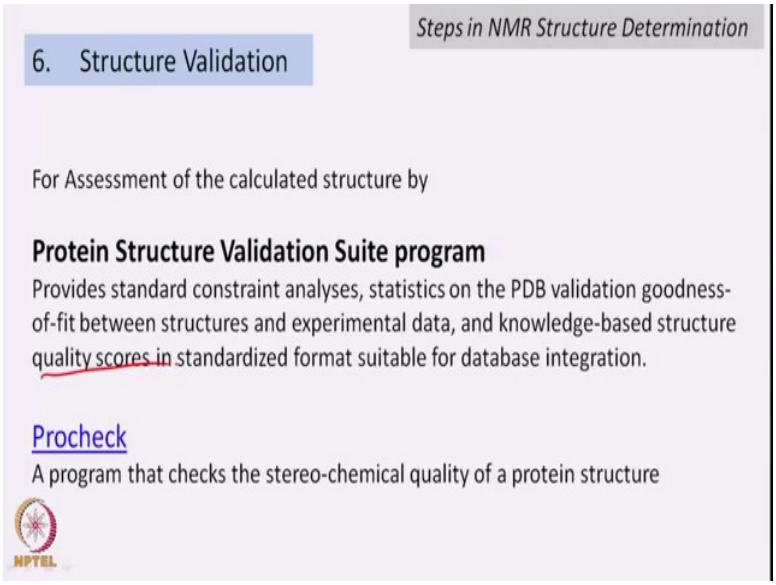
(Refer Slide Time: 28:25)



So in nutshell what we have done now, starting from the polypeptide chain using, we did MD simulation to calculate it, used the restraints for, force it to correct like contexts, then we get a 20 minimum energy structure and from there you can take one of the representative structure that will be a structure of your peptide.

If you look at here starting of the same 18-mers that we have started, there are some amino acids which shows a helical and some appear as a tail. So that is what the energy structure determination protocol based on the minimum energy is used for getting this kind of the structure.

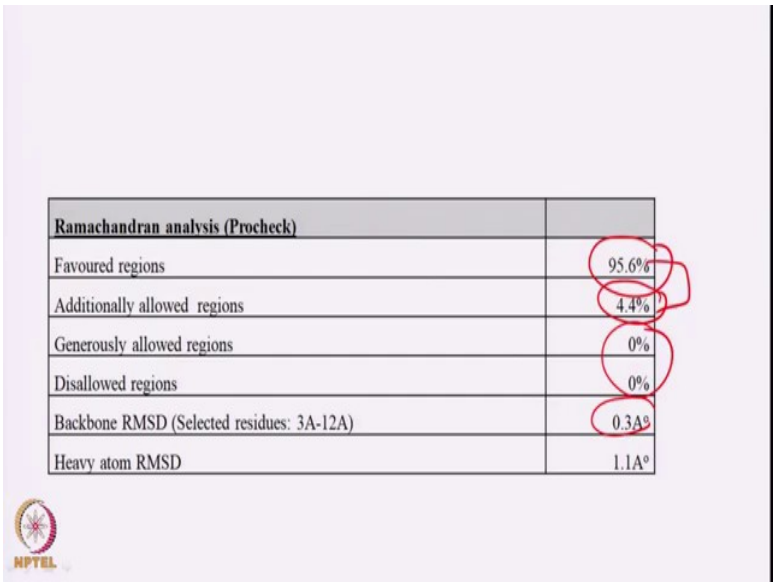
(Refer Slide Time: 29:12)



Now the next step is to validate this structure. So for assessment of the quality of the structure one has to use this protein structure validations suite. So essentially what you need to do to provide the standard constraints analysis and a statistics on PDB validation protocol and then these structures will be tested against already solved structures using these protocols and then it will, quality score will be given for your structure and that will be suitably integrated with the database.

So what we are doing, just for our obtained structure we are checking against the already solved structure from the PDB. And finally you need to check the stereochemical quality of your protein structure. That means you have to do the acid test for your solved structure by putting it into Ramachandran Plot map, whether all amino acids are falling into the Ramachandran Plot map or not.

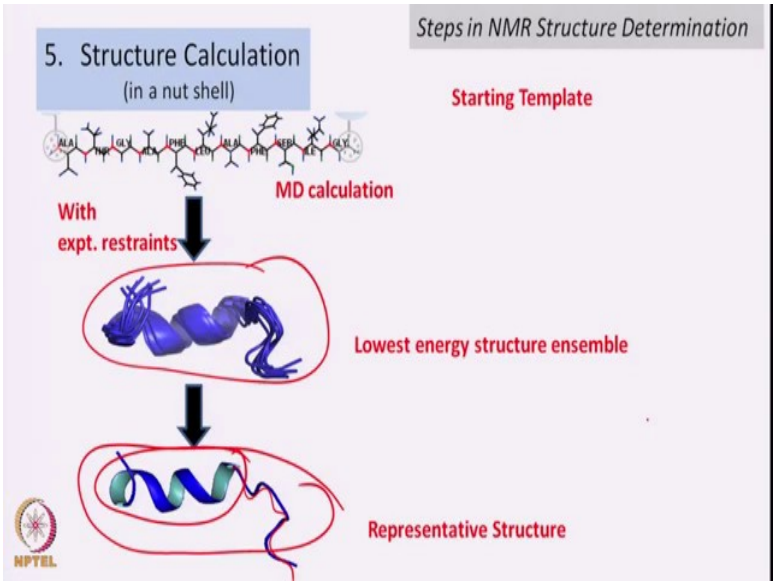
(Refer Slide Time: 30:11)



If you do that, so you can find it out like your this, some of these protocols we will find it out how many residue or what is the region in the favorable region of the Ramachandran plot? What is additionally allowed region? So if you look at the previously determined structure most of them are either in allowed region or additionally allowed region. None of them are essentially in disallowed region.

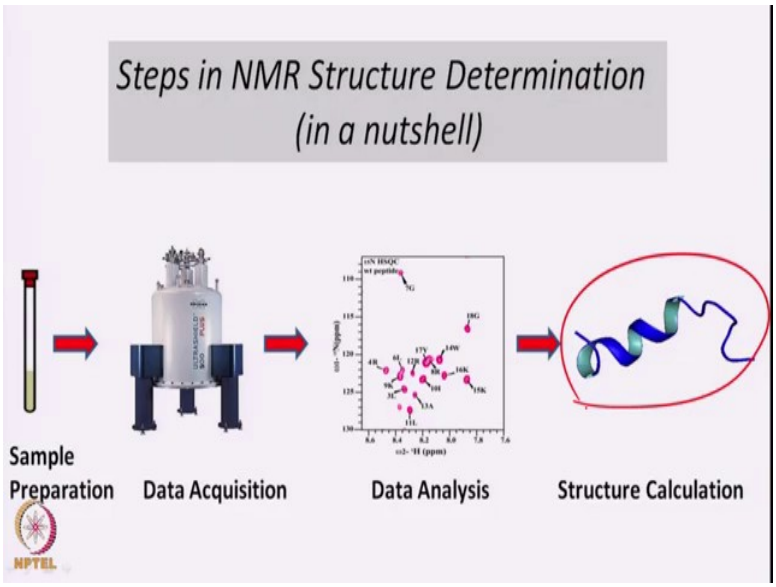
That means this structure is very good, so quality of this is very good. And rmsd is root mean square deviations, so deviations of C_α items of those 20 amino acids is very minimal. That means structure that has been solved is very, very coherent and you can see it from here.

(Refer Slide Time: 30:54)



All these structures seems to be overlapping on top of each other. These are 10 minimum energy structure, okay. So if you do that we have final structure determined.

(Refer Slide Time: 31:03)



So to sum up what we did here, we started with a sample, dissolved that in a proper liquid at 10 percent of D2O. Then we take it to high field magnet or whatever magnet that you have. Acquire data for days, then we sit together and do the data analysis for weeks and then we calculate the structure and that will be final representative structure of a peptide.

And that is what, and then we need to validate it, do the procheck analysis to see the structure is of high quality. If we obtain this we are done. We can submit the PDB and that will be final

structure determined from NMR spectroscopy using solution-state protocol. Okay so I hope I gave you glimpse of how you determine the peptide structure using NMR tools. I hope this will be useful for you and if you have any doubt, please do write to us, ask us. We will try to resolve them. Thank you very much.