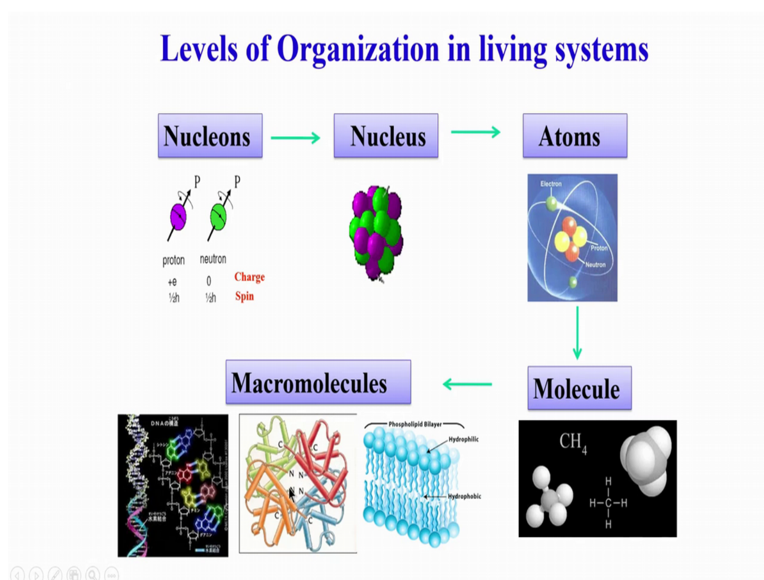
Multidimensional NMR Spectroscopy for Structural Studies of Biomolecules
Prof. Hanudatta S. Atreya
Department of Chemistry
Indian Institute of Science, Bangalore

Lecture - 21
Basic concepts of protein structure

So, we begin with the second part of the course from this lecture onwards. In this part we will look at how a protein structures are solved by NMR. How what kind of isotope labelling techniques are required, how 3D NMR experiments are used for assignments, and in the last part of the course we will look at how protein ligand interactions can be studied.

So, let us begin with very basic understanding of what is a protein structure, how does it look. What are the different aspects of a structure of proteins, so, that we can understand how NMR then can be used for solving them.

(Refer Slide Time: 01:09)



So, this is what is now we start from. So, we looked at from the very beginning how the nucleus is made up of nucleons and that is how the NMR began looking at the spin of the nucleus. And that is, what is the most basic element in an atom of course, this can further be broken down, but we look not go beyond this nucleons. So, given this is the most basic level of organisation that is most basic elements from which all nuclei in this

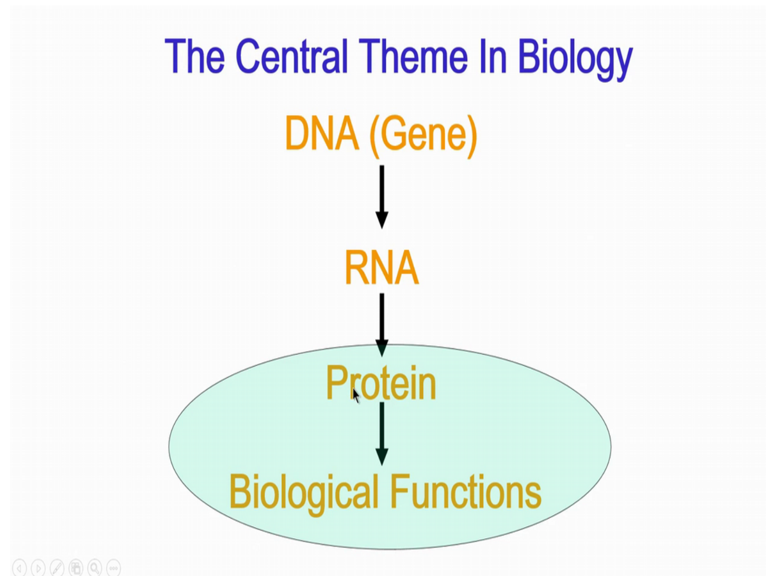
universe are built. Now, these nucleons are packed in nucleus it is a very tightly packed. And, there are lot of great force which holds them together and nucleus is located in an atom.

So, atoms are made up of nucleus and electrons which are circulating around the nucleus and that is concept of atoms. And, atoms actually organise or joined together to form molecules. This is something which all of us know, but what is important is beyond molecules. If you look at molecules which are made up of atom molecules themselves can arrange a self-assemble or organise in different hierarchical structures. They can go in to different complicated complex structures either by self-driven or it can be through induced assembly.

So, these higher order structures which are formed from the molecules are known as macromolecules. So, macromolecules is the world basically for molecules such as nucleic acids DNA, proteins, or membranes, or lipids by layers. So, these are complex molecules. So, we look at how proteins can be studied by NMR, because other two are very complicated on their own they required altogether different topics to be covered. So, we look at protein, but down the line we can also look at how protein interact with this different components such as nucleic acid and a membranes. So, this is just a schematic of how the different levels of organisation in a living system.

So, we are looking at atoms actually nuclei by NMR which are located in the atoms of these big molecules. So, we will be probing by NMR each every nucleus here in this molecule. So, by probing every nucleus in this molecule NMR is a very unique technique which allows us to get very high resolution information. High resolution meaning we can get atomic at level details of each and every atom in this molecule, because each and every atom will have a nucleus and that nucleus can be studied by NMR. So, that is the unique aspect of NMR that it gives us high resolution information or structural details into molecules or being molecules.

(Refer Slide Time: 04:01)



So, this is the central dogma in biology which again is very well known, that at a very basic level you have a genome that is DNA which is there in the nucleus, and that is transcribed into RNA. That is this transcribed into RNA and there are different types of RNA, but finally, this RNA messenger RNA which codes for a protein is translated or converted into a protein in the cell. And, that protein then which is produced carries out the different biological functions.

So, our focus is going to be on this aspect, how protein carry out the different biological functions.

(Refer Slide Time: 04:43)

The study of Gene/Protein
Structure and Function at a
Molecular level
is called STRUCTURAL BIOLOGY

So, this field where we study the structure of protein or it is interaction with the gene or how does a gene function or a protein function with respect to structure at a molecular level; that means, at an atomic level that field we call it as structural biology. So, structural biology is a very emerging area where we look at different aspects of how a protein structure correlates with it is function and how does it interact with different ligands in the system.

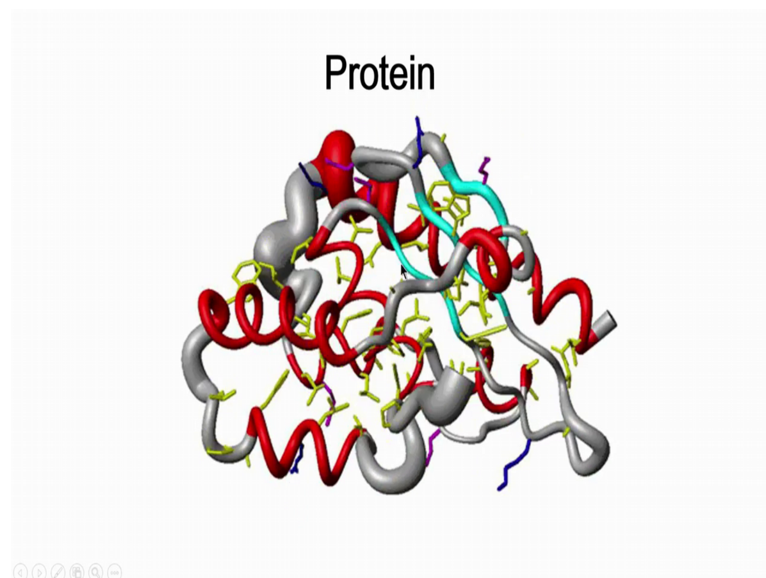
(Refer Slide Time: 05:15)

Some Proteins.......Have a
Specific 3D Structure

So, these are some basic aspects what we are covering now as we go down the line we

look at some more details. So, for example, now some proteins not all of them again we will now look at proteins which are different from this. So, some proteins in in universe in our body in a cell or an organism, they have a specific 3-dimensional structure. This is very important to bear in mind, because proteins are actually not just simple molecules in our molecules they are dimensional molecules complicated or macromolecules, where we have to understand how the structure is formed how it works in space. So, this is a schematic again of some protein which is shown in a animated mode.

(Refer Slide Time: 05:59)



So, you can see here the red colour turns which are seen here are the helices alpha helix. And, the blue colour light blue cyan colour things are the beta strands or beta sheets and the grey colour here are the loops. And, you can see several side chains of amino acids shown in yellow this is the stick model of amino acid. So, for example, this you can see here is an aromatic amino acid hydrophobic hydrophilic and so on.

So, one of the special feature of protein that is a globular protein which occurs inside the cell or outside the cell is that the hydrophobic groups are located inside the protein structure. So, we can see inside this it is buried in the core of the protein are the hydrophobic amino acid such as leucine, isoleucine, methionine, phenylalanine, alanine, and so on. Whereas, on the surface you will have more hydrophilic amino acids like this blue colour here is a lysine glutamine aspartic acid as per gene and so on.

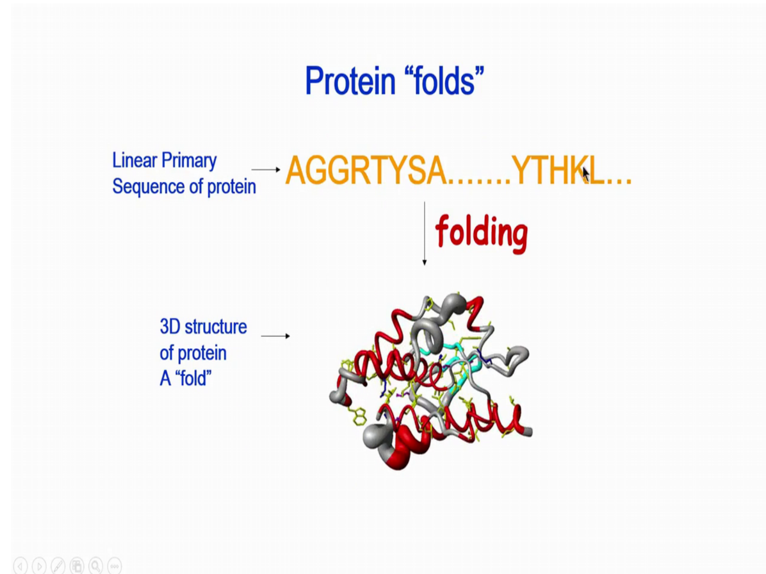
So, that is an interesting the part in a protein that the inside is hydrophobic outside is

hydrophilic, but only happens the that can happen only in a globular when it is inside a cytosol or outside the cell not in the membrane case. So, will membrane protein are actually opposite. There the hydrophilic groups are located outside the surface on the surface whereas, the hydrophilic are inside. So, the hydrophobic is outside hydrophilic is inside is exactly opposite of what you see in a globular cytosolic protein. And, why is that because the membrane environment which we saw in this slide here, let me go back you can see here the membrane is actually hydrophobic in nature. So, the amino acids which are facing this bilayer they should actually be hydrophobic, because hydrophilic cannot attract attracted to hydrophobic groups.

So, hydrophobic interactions make the hydrophobic part of the protein interact with this that is outside of the protein whereas, inside the protein hydrophilic groups are then located. Now, this grey colour you see there is again a slight difference in the thickness this is very thick here then this thins out, what is that indicate that is a schematic again to indicate the dynamics of the protein. So, what you see here on the screen we are only seeing a rotating protein structure, we are not seeing how the different dynamics groups are dynamic with respect to each other. Because a protein actually breaths, it fluctuates, all the bonds are moving and fluctuating at different time scales. That kind of a dynamics is not captured in this picture, but what is being captured is that how the loops are having the higher dynamics.

So, the thickness here indicates that the loops are having more dynamics means this this loops for example, here can span wider range of confirmation and therefore, it is thicker compared to this helices here which do not span a large range of confirmation their rigid. So, in their rigid their dynamics is less, but loops which are between the helices and beta strands they are always very dynamic because they do not have any particular structure. So, this is just general again introduction to protein structure, which you can refer to in many of the books and which shown in the beginning and let us continue with this look at some more basic aspects of protein structure.

(Refer Slide Time: 09:51)



So, now this protein is made up of a linear sequence of amino acid. So, you can see this is a code letter code single letter code for alanine G is for glycine, R is for arginine. So, each of the 20 amino acids have a single letter code assigned to them. So, for example, Y is for tyrosine ok. So, there is no direct correlation with the first letter because tyrosine has T, but T is something else this is threonine. So, I would recommend very strongly to actually memorise that what is the letter code stands for because when we go down the line when we look at protein assignments, we will not be again referring to the we will be referring mainly to the single letter codes.

So, if I say immediately RTY one should be able to recollect that this stands for arginine threonine tyrosine; so, there each amino acid as therefore, a single letter code. So, now you can look at the sequence of a protein this is called the primary sequence set or primary structure of a protein, it consist of linear sequence of amino acid linear there is no branching here. This could be in the number of amino acid depends on the size of the protein. For example, if it is hundred amino acid protein it will be about 100 in to 100 13000 molecular weight. So, the molecular weight will become around 10 Kilo Daltons. So, this is come based on the fact that an average amino acid a weight of average molecular mass of an amino acid is 100 100 and 10 Daltons.

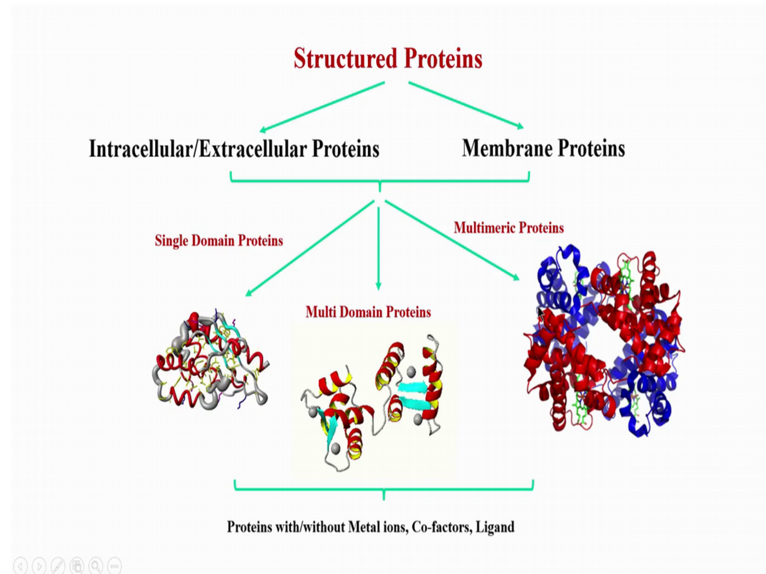
So, now this is linear sequence of amino acids if it is allowed to fold it starts folding into a 3 dimensional structure, but it first actually undergoes also a folding to secondary

structure which is not shown here, a secondary structure is basically alpha helix which is shown here beta sheet. So, they start forming first and then these alpha helix and beta sheet come together and then they form the final 3-dimensional structure. So, this is the concept of protein folding ok. And, this happens very fast this from sequence as soon as the protein is coming out from the ribosome it starts folding like this. So, it is happens at the scales of milliseconds sometime very fast can be microseconds, but typical millisecond is a timescale for protein folding. So, that is very fast from NMR point of view.

So, NMR actually what we do is we do not study this process of folding very often in a real time. We study the final 3 dimensional structure which is already formed. We can study the intermediate structures from here to here when you go you can study some kind of intermediate structure, if you can arrest or you can hold that structure at different time points. There is something which we will not do in this look into this course there is the large area of research, where people have tried to track, how the sequence of amino acids linear sequence gets folded into a 3 dimensional well defined structure ok.

So, this is the protein folding problem, but what we do in NMR is we will look at the final structure; that means, we want to solve the structure of this part. I mean we want to see how this looks. We will have this information with us and using different NMR experiments, we will have our goal is to get this structure. So, that is our goal in this whole concept of protein structure determination.

(Refer Slide Time: 13:21)



Now, proteins are actually we divided into different categories. As, I said earlier one is one category we can say globular that is intracellular or extracellular protein. This means it is inside the cell or it is outside the cell and once my protein which are sitting in between that is on the membranes, they are called as membrane proteins. So, there is a fundamental difference between the 2 as I said here the hydrophobic groups of amino acids are inside the core of the protein, and the surface of the protein is made up of hydrophilic amino acids. Whereas, here it is exactly opposite we have hydrophobic amino acid outside and hydrophilic amino acid are inside the core of the protein.

Now, these proteins both of them can be further categorised into 3 categories. Number one they can exist as a single domain protein. A single domain is an example I just gave in the previous slide, where the entire chain folds into a single structure ok. So, this is called as a independent folding unit. So, if you take the sequence which e of this protein, it will form as 3-dimensional specific structure, but it will remain alone means it will not associate with any other structure, it will be a single chain structure. But, you can also have what is called multi domain proteins. So, if you look at this here example a multi domain means this is one protein you can see independent, this is another independent protein and they are actually joint.

But, there is a single chain again meaning it starts from one end of this protein, let us say here and goes all the way up to the end of this protein which is somewhere here. So, this

is a single chain means it is a single chain, but it has got divided into 2 parts and these 2 parts are linked by a linker. So, this is a linker this loop here which you can see in grey colour is a flexible linker meaning it is flexible. So, this domain and this domain are actually independently moving, they would not move together this will have a separate moment this will have a separate moment.

But, they are actually also independently folding meaning, if I take this only this part cut it out here. Suppose I cut out this protein here and I take only this part of the protein it will get the same structure independent of whether this was present or not. Similarly, if I cut out this here again and take this portion put it in water, it will get this now out of this structure which is independent means, it is does not depend on whether this part was available or there in the protein or not. So, therefore, they are kind of independent folding units.

Then you can have another type of structure where different units that is this for example, take the single domain protein multiple units of single domain protein they associate with each other and form what is called a multimeric protein. So, multimeric proteins are those which have multimers, trimer, tetramer, dimer and so on. So, you can see this is the example of the famous haemoglobin molecule which is at a tamer in the functional form ok.

So, this is how many proteins work they do not have a single unit they have either multi domain or 2 or 3 or more and or they can actually even associate with more of the. So, here you can see they are not single chain. This red and blue are actually independent, they are separate molecules, they can be easily broken into separate force separate molecules using some treatment. So, they are not covalently linked.

So, if you look at this picture again here there is a covalent link between the protein this domain and this domain they are not lose separated, but here they are not covalently link they are noncovalently associated, because of some sort of interactions and therefore, they can be dissociated also ok. So, these are the different categories and we will see that for different types of proteins we need different types of strategies in NMR to solve the structure. Now, further or each of this now can further be associated with metal ions for example, you can see here I shown in grey colour these are metal ions.

Now, this metal ions are because of the requirement for the protein to function, it may be

a protein which needs a metal ion for function for catalysis and enzyme or it may be that it is storing those metal ions. In this case this is example of a calcium binding protein. So, this grey colour balls here are actually calcium ions. So, they are this protein is known to store calcium when the calcium come becomes excess in the cell, but this is one type of metal ion, you can also have different types and this metal ions can be again categorising to diamagnetic and paramagnetic.

Paramagnetic metal ions are those which have an unparalleled electrons and that has a lot of implications in NMR, because paramagnetic cause ions, metal ions, they cause very high relaxation very fast relaxation of protons which are surrounding this. For example, let us say assume that this is a metal ion which is paramagnetic all the protons which are in the zone here will start getting broadened. Broadened means signal will start disappearing, because of this interaction of paramagnetic dipole moment with the nuclear dipole moment electron and nuclear interaction.

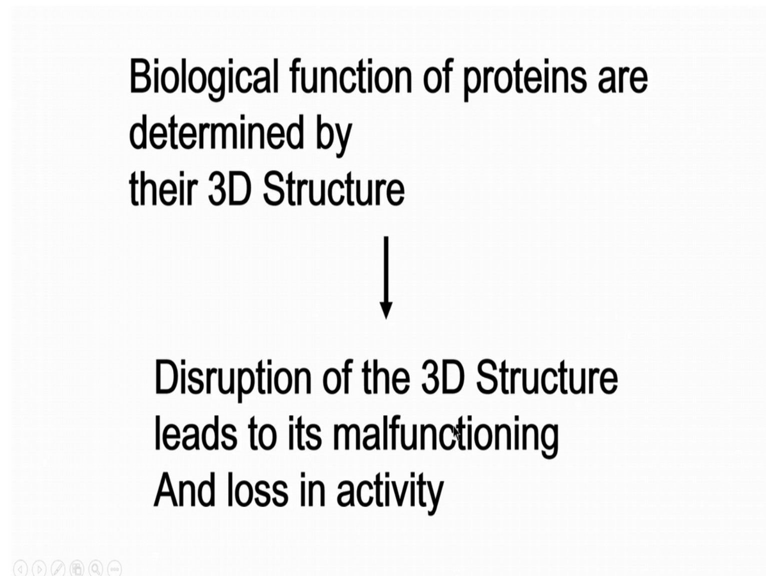
So, that becomes very challenging for NMR to be used because broadening of lines means sensitivity goes down. So, therefore, paramagnetic metal ion protein structure are very challenging part and has been dealt with by several groups, and that is again we will not go specifically into how to solve paramagnetic protein structure in this course, but you should remember that this metal ions are also part of the protein they are required for the function of the protein. So, when we solve a structure we should be aware whether a metal ion is there in the protein or not. There are varieties of techniques to find that out we will you can refer to any spectroscopy method and that can be uses.

Now, we could also have cofactors for example; look at this here haemoglobin as a heme, the porphyrin ring as a cofactor. So, these are also we required because haemoglobins main job function is for transporting oxygen. So, oxygen has to bind and there is an iron here. So, that is metal ion again paramagnetic and this porphyrin ring is holding that metal ion, and it has very important role to play and that is like a that is called a cofactor.

Now, the other option other thing is ligand say sometimes the protein interacts with DNA or protein can interact with another protein and then only it will function. So, again that always in a cell it may not exist like this alone, it may exist in a form in which it is bound to the ligand. Ligand can be any small molecule metabolite or it can be small part of

DNA, RNA etcetera or a peptine so on. So, ligands are also part of the protein in the sense that they are required for the protein to works. So, the native protein always exist bound to the ligand in the cell. So, these are the other next level of complications that is based on the each protein can further have these things available or present in the structure; so, this when the basics of protein structures.

(Refer Slide Time: 20:59)



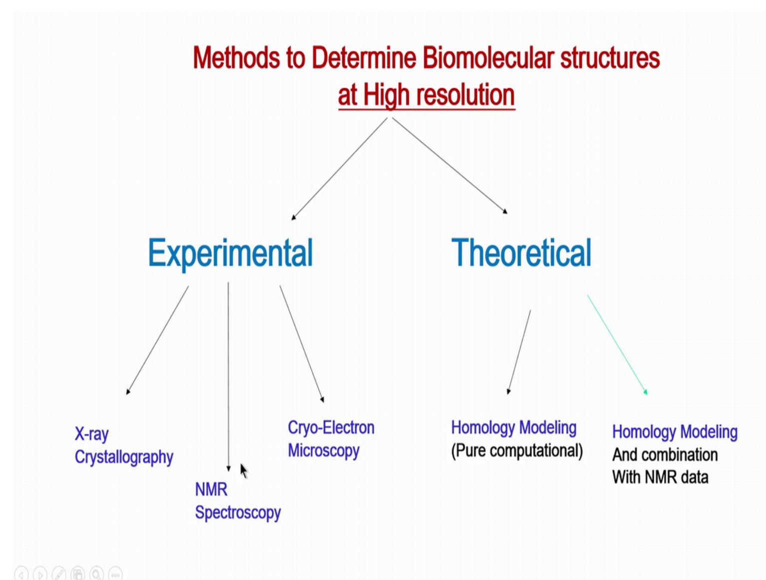
Now, let us see what happens to the protein structure goes deformed. So, for example, we know this this is a very broad picture that the biological function that is what kind of function it has in a cell depends very much on it is 3D structure. So, for a proteins which have 3 dimensional structure. So, if I disrupt the 3D structure of the protein for some reason for example, it can be mutagenesis amino acid in a protein can get mutated to some other amino acid, or a range of a protein can amino acids can get mutated or the protein may get denatured, because of some acidic or alkaline conditions in such cases the 3D structure is lost.

So, when the 3 structure is lost whatever function it was carrying out now cannot be carried out. So, that is called as loss in activity. And, this results in various problems in the cell or in the overall organism. For example, if you have Alzheimer’s disease there you have a protein which is not properly form the 3 structure it gets miss folded, and start forming amyloid fibrils. And, that fibrils can lead to plucks in the brain which cause the problem.

Similarly, many other protein sickle cell anaemia is a well-known disease well known problem in which one amino acid in haemoglobin is mutated, and that results in a change in the structure of the protein and loss in reduction in activity. So, like that there are many possible changes in structure happening in our body, which can result in the structure getting disrupted or basically damaged and that can result in the malfunctioning. So, therefore, the goal of NMR and drug discovery and so on is to find out what is the, what makes the 3D structure go bad, or how we can retain the 3D structure by interacting with some ligand. Sometimes some ligand if you add to the protein structure it gets stabilised.

So, can we study that kind of interaction and stabilise the protein structure. So, 3D structure is therefore, very important to study for proteins because their function or their actual role in our cell depends on that structure.

(Refer Slide Time: 23:19)



So, how do we determine 3D structures of proteins? There are 2 main categories; one is called experimental method or you can use completely theoretical approach. So, in experimental approaches the well-known methods are X-ray Crystallography, which is a very old technique and this is now well established and a lot of protein structures are now being solved with X-ray and second approach is what we are going to look at it is NMR spectroscopy. In one difference between X-ray and NMR important difference is in X-ray you need the crystals of the protein. So, many proteins do not crystallise.

Therefore, those kind of proteins you have to do with NMR spectroscopy only you cannot use X-ray.

But, another the disadvantage of NMR is a size typically as a protein goes beyond 50 kilo Daltons it becomes very difficult to solve the structure easily. In fact, up to 50 40 to 50 also is a higher range for NMR spectroscopy to solve to get high resolution structure. So, remember our goal here is high resolution meaning you want every atom, or every amino acid in the protein to be resolvable, or separable, or we should know the coordinates of all of them. It is not just we want to know whether it is having alpha helix and beta sheet and so on we want to actually know the 3 dimensional structure.

So, for NMR the limitation comes from size higher the size of the protein that is beyond 50 KD it is very difficult even with modern techniques. So at crystallography does not have that limitation, but for crystallography you have to crystallise and many times many proteins do not crystallise. And therefore, that is a difference advantages that basically complementary methods.

We will see shortly that there is another class of proteins, which again cannot be solved with crystallography only NMR can help us there we look at that in few slides from now. A, experimental technique another which has come up recently as becoming popular is cryo electron microscopy, again this requires the freezing of the molecule like here, freezing meaning crystallisation or cryo which is frozen.

So, again this is looking at a static structure. So, that is a difference between these 2 an NMR and NMR you are looking at a protein, in the solution state meaning the protein is in it is native form because in our cell the protein is always along with water. So, we are looking at in a native form in NMR in solution NMR spectroscopy.

So, therefore, we are looking at the native form of the protein and the complete dynamics of the protein is also available to us whereas, in these 2 techniques the molecules are frozen. So, in that sense we are getting a snapshot not a dynamics we want get the dynamics of the protein we get a snapshot of the protein structure. So, that is a difference whereas, NMR helps us, but again as I said NMR NMR has it is own limitations. So, one has to use this in a complementary manner.

Now, one can use a purely computational approach. You do not have to actually do any

experiment, do not have to make any sample, you do not need all these instruments, you can simply sit on a computer and get the structure of a protein again in very high resolution. And, that is using a homology modelling concept. Homology modelling means, you have another protein which has a similar looking which similar identical sequence, remember we are looking at primary sequence. We do not have the structure, but with the primary sequence is very similar or identical to another protein for which the structure is known.

So, let us take a protein A and protein B, if protein A you have sequence matches with protein B to some extent and what is that extents at least more than 30 percent identity. Means at 30 percents locations or more the protein A and protein B have identical amino acid. So, such a case if protein b structure is known then protein a structure can be solved by simply going to computer. You do not need to really take protein a and go through all these experiments. But of course, the real reliability of homology modelling goes down, if the identity is less than 30 percent.

So, if it is 90 percent or 95 percent identity 2 protein A and B are 95 percent let us say identical, means only in 5 percent of their residues are different, then most likely they will always have the same structure. So, that can be now determined with just a computer. That is saves lots of time and that also increases the speed because this is faster compared to these techniques.

Now, there is another option we can do which has again come up in recent days is that you take homology modelling, which is already the taking the known structures and sequence, but you combine with some NMR data. Let us say you have some preliminary NMR data of that protein A and B, suppose we record for let us say we are looking at protein A and B and protein A is known structure ok, and protein B is unknown and protein B has having a very high sequence identity with protein A.

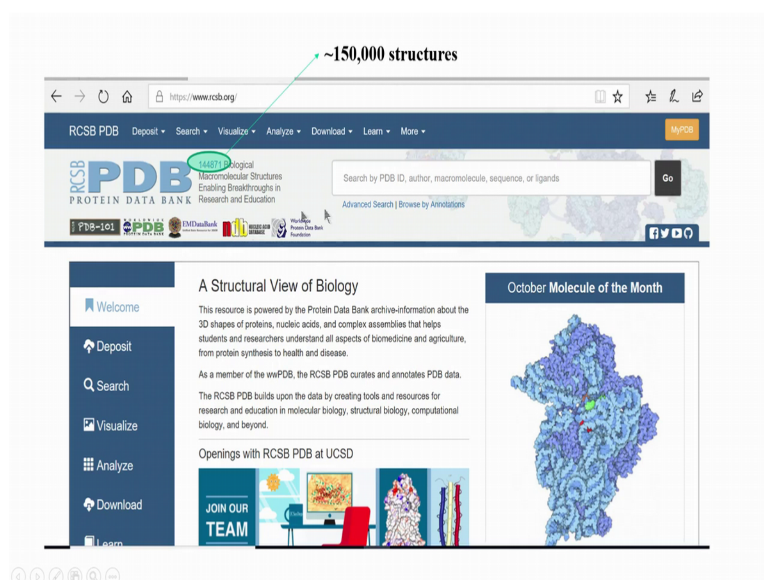
So, homology modelling can be done like here, but additionally I can take some more information recording data on protein B ok. So, if you are record NMR data of protein B and get some proton carbon experiments and spectra and chemical shifts, I can use that information and further improve my modelling.

So, that helps to improve the modelling and that is the combination of these 2 homology modelling and NMR data. So, this is like a semi is a combination of experimental and

theoretical. So, it is not purely theoretical approach it uses the chemical shifts knowledge which comes from experimental data.

So, these are broadly the different techniques, which are nowadays available to solve protein structures, but actually what do you do with this protein structures. Once you have a protein structure how what do you do with that to you. So, typically these have to be deposited in a data bank ok.

(Refer Slide Time: 29:37)



So, that is the most protein most popular data bank is called the protein data bank. In a protein data bank the protein structures are deposited and you can see as of now that is just recently there are about 150, 000 nearly 150, 000 structures in this data bank. So, I would recommend very strongly to you to visit this website which is just shown here and actually have a look at this. What are the different types of information structures available, because this is where the complete the history of data of protein structures are available. And, this is where you have to deposit all the protein structure before you publish.

So, if you have a new protein and you have solved NMR or crystallography or so on you need to deposit here and get it verified, whether that protein is structure is correct or not by different techniques and then only you can go ahead and publish. So, in the next class we will continue with the basic protein structure and we will see what are the different types of protein, there is another class which we have not looked at we look at that and

then continue with how NMR can be used for protein structure determination.